
NATURE INSIGHT RNA SILENCING
nature
ANTARCTIC WARMING
Climate reconstruction gets to the heart of the continent
WHO DO YOU THINK YOU ARE?
Personal genomics changes the rules
SOLAR SYSTEM EXPLORATION
The Titan-versus-Europa dilemma
SEXUAL REPRODUCTION
A long wait for Aspergillus
NATUREJOBS
Biotech & pharmaceuticals

nature

www.nature.com/nature

Vol 457 | Issue no. 7228 | 22 January 2009

China's wind-power potential

The nation can lead the world in wind energy — but its policies need to be more coherent.

A quick glance at China's wind-energy statistics suggests that all the right things are happening. The country has doubled its capacity every year for the past three years; in 2007, it had surpassed a 5-gigawatt target three years ahead of schedule and, in 2008, it hit a revised 10-gigawatt target two years early. Domestic manufacturers are positioned to produce more wind turbines than any other country over the next three years.

But the reality is much more complicated. Many wind projects do not even get off the ground because power companies cannot make enough money from them. Those that do go forward often produce turbines that simply sit, waiting for four months or more — a big delay in financial terms — to be hooked up to the electricity grid. And even when they are connected, they break down more often and are much less efficient at producing energy than those in many other countries (see page 372).

In May 2007, the Global Wind Energy Council, the Chinese Renewable Energy Industries Association, and the China Wind Energy Association released a joint memorandum proposing improvements in the industry. More than a year and a half later, the concerns expressed in the memorandum have been borne out and the recommendations remain just as pertinent.

For example, the bidding system used by the government to appoint developers favours companies that agree to supply electricity at cheaper prices — even if that price will render them unprofitable. As a result, many projects haven't even got started. International development companies with more experience and foreign turbine-makers with more efficient machines don't even bother to bid. The memorandum recommends setting 'feed-in tariffs', which offer a guaranteed rate for power supplied and offer developers more consistency and planning. China has moved in this direction, but the process of allocating the projects is still opaque and, from the perspective of developers and turbine-makers, frustrating.

The country also needs to significantly improve its grid, and to coordinate it with renewable-energy developments. Grid companies are understandably not keen to embrace energy produced by wind — an erratic and relatively expensive source — so it will take incentives to make the grid companies want to play ball.

Another way that China has discouraged foreign developers from entering the fray is by preventing companies with less than 51% Chinese ownership from taking advantage of the Clean Development Mechanism, which allows developed countries to offset their carbon-reduction commitments under the Kyoto Protocol by investing in sustainable-energy projects in developing countries. If one has to be protectionist, surely it is better to do so in a way that bolsters domestic companies rather than simply penalizing foreign ones? Although some of China's policies have used this approach — it requires, for instance, that 70% of turbine parts be produced locally, encouraging foreign companies to build manufacturing facilities in China — it would do well to extend this across the board.

"China needs to significantly improve its grid, and to coordinate it with renewable-energy developments."

China could still learn a lot about turbine manufacturing and wind-farm maintenance and management from countries that have much more experience in wind energy. One place to start would be to end its obsession with the number and capacity of its turbines and focus instead on producing power from them.

The speed with which the nation has scaled up to 10 gigawatts of wind energy is impressive. But if China could harness the 3,000 or so gigawatts of wind estimated to be available in the country, it would be able to cover almost all of its current electricity demand. That figure won't be achieved any time soon. But China should make good on plans to hit a more reasonable target — upping its 2020 projection from 30 gigawatts to 100 gigawatts. Operating at international standards of efficiency could produce 5% of the nation's energy needs and, depending on US policy over the next few years, make China the biggest producer of wind energy in the world. But only if China takes a more aggressive and rational approach will it make the most of its wind. ■

Dismal no more

Europe's Joint Research Centre should be empowered to stimulate other EU institutions.

The European Union is run by its council of ministers and its parliament, with the European Commission as its executive body. But there is another component that is generally unloved and yet has a crucial role: the Joint Research Centre (JRC). The seven large institutes that make up the JRC have remits that range from energy and environment to health and security. With a budget that approaches €400 million (US$520 million) per year, the JRC is responsible for providing the scientific and technical support for EU policy. And according to a fairly positive evaluation released last week, led by David King, former science adviser to the British government, it now does this quite well.

The dismal reputation of the JRC originated in 1980s, after changing priorities in Europe had forced the centre to shift from its original mandate of researching nuclear energy, safety and security. It reinvented itself, and diversified its research fields, in a largely undirected and unmonitored way. Politicians soon started to complain about its lack of a clear mission and its inefficiency. In 1998, the institutes' labs were relaunched with a new, customer-orientated mission, tightly harnessing them to EU policy support and fulfilment.

The various JRC centres are now reasonably efficient machines,

pretty well stripped of the dead wood that had been dragging them down. They fulfil just about every task their customers — that is, the EU bodies — set them, often with great success. The King report acknowledges, for example, that the biotechnology division has become the world leader in setting standards for detecting and monitoring genetically modified organisms in agriculture and food, while also playing a key part in the implementation of the controversial new Registration, Evaluation, Authorisation and Restriction of Chemical substances (REACH) regulations.

In the process of becoming customer orientated, however, the JRC lost its independent research activities. Does this matter, given that they were not highly regarded anyway?

Yes, it matters. The role of science in policy has become ever more important. Crucial decisions about climate, energy and the like rely on the highest quality of scientific information. JRC scientists need to be able to interact intelligently with cutting-edge research in the academic community, and to extend it appropriately.

But high-level scientists need an appropriate intellectual environment to work in, and this is where the JRC has had its hands tied. Right now, JRC researchers are running to stand still. As new tasks stream in from their demanding customers, research projects enthusiastically started during a previous task are dropped — even if they might have had long-term value for policy. The scientists tend not to publish much in academic journals, as this has not been considered one of their 'deliverables'. Most JRC science is written up in internal reports. These are not ideal conditions for attracting and keeping top, ambitious scientists.

This is why the King report's proposal that a small proportion of the budget be directed into exploratory work is welcome, as are two other radical proposals, to some extent related.

First, the JRC needs to develop its own long-term strategies so that it can plan more rationally. Although it has 2,750 staff, the scatter-gun tasks required of it spreads them too thinly. So the strategies should also set out criteria for which tasks to accept so that critical mass can be achieved from limited human resources.

Second, the JRC needs more political responsibility. Given its accumulated knowledge and experience, and its close contact with the scientific community, it is well placed to see problems coming. Structures should be set up to allow it to advise, not just serve, its customers.

In short, the JRC needs to be allowed to grow up. Europe's council, parliament and commission now need to shed their distrust and make the JRC an even more useful institution than it has already become. ■

Choosing a world

Titan is a slightly more appealing lunar target than Europa for the next outer-planets mission.

In just five decades, planetary spacecraft have provided an extraordinary wealth of discoveries: the oddly young surface of Venus, the ancient landscapes of Mars, the volcanoes of Io, the geysers of Triton, the lakes of Titan, the ocean of Europa. But two of the most sought-after things have not been discovered. One is life. The other is how to explore space cheaply.

The eye-watering expense of major missions imposes on the planetary enterprise a glacial tempo that is at odds with the drama of its discoveries. This is particularly so the further they get from Earth. The United States is currently deciding whether it should send its next 'flagship' outer-planets mission to Europa, an ice-covered moon of Jupiter, or Titan, the thick-atmosphered moon of Saturn (see page 366). The decision — in which Europe, as a partner in the mission, also has a stake — will shape the lives and interests of researchers through to the 2040s. But with total costs of more than US$3 billion, only one mission can be afforded.

Until recently, Europa was widely tipped as the favourite. Its ice-covered ocean was one of the two most promising sites the Solar System offers for life beyond Earth. Indeed, it might have greater attractions than the other prime location, the subsurface of Mars; given that impacts can send meteorites flying between Mars and Earth, there is a real chance that any Martian life found would be related to that on Earth. Europa, which is much more isolated, is a better bet for an independent origin of life.

Unfortunately, to do full justice to such possibilities would require measurements from Europa's surface, or even below it — measurements that technology cannot yet offer. Instead, the proposed Europa mission would be an orbiter restricted to producing images in various wavelengths, including those from ice-penetrating radar. The data would bring some closure to arguments about the thickness of the crust (see page 384) and might identify places where the subsurface can be accessed most easily. But in the end, an orbiter would be only a precursor to the next act — a landing mission, perhaps with drilling capability — the *dénouement* of which may be half a century away.

A Titan mission, by contrast, would be unlikely to encounter life, but would be much more intimately involved with its environment. It would include a Titan orbiter with radar and other instruments to map the moon far more thoroughly than the Cassini mission's ongoing fly-bys can; a European lander designed to float on one of the hydrocarbon lakes; and a hot-air balloon (or more accurately, a slightly-less-cold-air balloon) that would drift around the moon studying its hazy atmosphere and rich landscapes of channels, lakes and dunes.

"It is hard to do science when new data come in a splurge every few decades."

The Titan mission would also offer technological firsts, including the floating lander and that undeniably romantic hot-air balloon. Although floating on a lake is a skill for which there seems to be no further call in this Solar System, ballooning could be used to explore many other planets and their moons, and developing this capability could be seen as an investment in that future.

It is hard to do science when new data come in a splurge every few decades. Gifted people go elsewhere, those who remain get locked in irresolvable debate, hypotheses calcify into dogma. But at the moment there seems to be little that can be done to break the long, drawn-out rhythm of feast and famine. And if we must choose, then floating below the rings of Saturn seems marginally more appetizing than above the hard ice of Europa. ■

RESEARCH HIGHLIGHTS

NANOMATERIALS

Squid suckers scrutinized

Adv. Mater. doi:10.1002/adma.200801197 (2009)
The fearsome suckers with which the Humboldt squid (*Dosidicus gigas*) clutches its prey are lined with toothed rings. David Kisailus at the University of California, Riverside, Henrik Birkedal at the University of Aarhus in Denmark and their colleagues have peered into the jaws of these suckers to find out what makes them so strong.

The sucker rings have sharp teeth made from parallel tubes that are hollow near the circular base and filled at the sharp end. This channel-like structure directly affects the mechanical properties of the sucker rings, the authors say, adding extra grip and shear strength for terrorising prey.

Surprisingly, the sucker rings don't contain chitin, usually present in the shells of crustaceans. The sucker rings' major amino acids are glycine, tyrosine and histidine. The authors propose that hydrogen bonds between histidine residues give the rings their rigidity.

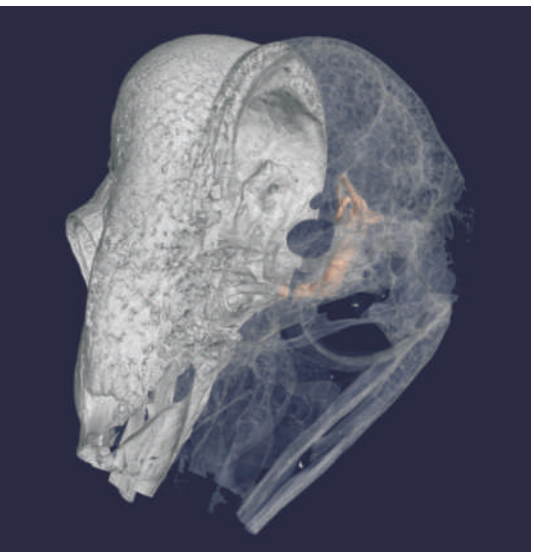
QUANTUM LOGIC

Blockade boost

Nature Phys. doi:10.1038/nphys1183 and doi:10.1038/nphys1178 (2009)
The groups of Antoine Browaeys of the University of Paris-South in Palaiseau and Mark Saffman of the University of Wisconsin, Madison, have independently demonstrated a simple set-up that could form the basis of a quantum logic gate, a system that has the advantage of having components widely separated enough to be easily addressable one at a time.

Their Rydberg blockades hold rubidium atoms in optical traps several micrometres apart. Lasers are used to excite an electron in one atom to an extent that blocks its neighbour from achieving a similar excited state. The next step will be to demonstrate useful entanglement by creating a working logic gate with the blockade.

NHM



CSIC–UNIV.GRANADA

Getting their morph on

Science **323**, 362–365 (2009)
Some precipitates take regular, sinuous and oddly lifelike forms, and are known as biomorphs. Juan Manuel García-Ruiz of the University of Granada, Spain, and his colleagues describe an intriguing chemical feedback mechanism that creates the microcrystals responsible. In biomorphs built from barium carbonate crystals, the formation of these elongated crystals has the effect of locally reducing pH, which allows the precipitation of silica onto the crystals, halting their growth and defining their shape. The properties of these microcrystals, which the authors observed using time-lapse video microscopy and electron microscopy, are responsible for the smooth curves and furled edges of the biomorphs.

The work opens the way for new approaches to the synthesis of biological and biomimetic materials, and to the exclusion of false positives when looking for life-like forms in poorly characterized environments.

PALAEONTOLOGY

Herd of hearing

Proc. R. Soc. B doi:10.1098/rspb.2008.1390 (2009)
If you want to know what *Archaeopteryx* sounded like, a first step is to work out what it could hear.

Working with specimens of 59 extant species of reptile and bird (including the barn owl, skull pictured left), Paul Barrett of London's Natural History Museum and his colleagues measured the length of a duct found in the bony part of the inner ear. They show that this length correlates with the hearing range and best hearing frequency of the animals, and can also be used as a guide to the complexity of their calls.

This relationship allows the first quantitative assessment of the hearing ability of extinct species: *Archaeopteryx* seems to have been a fair match for the emu in this regard. More data may allow inferences to be made about the bird's vocalizations, and thus reveal what sort of social structure it lived in.

NEUROLOGY

Serotonin and social anxiety

PLoS One **4**, e4156 (2009)
Certain versions of a key gene for the brain regulator serotonin disproportionately predispose those with a history of childhood abuse to depression and alcoholism.

Rhesus macaques and humans have similar variation in the serotonin transporter gene. Work by Karli Watson of Duke University in Durham, North Carolina and two colleagues shows that the animals show similar behaviour if they possess one 'short' version of the gene. These macaques demonstrated the simian equivalent of social anxiety by avoiding looking at pictures of faces, and particularly avoiding looking them in the eye.

PLANETARY SCIENCE

Martian methane

Science 10.1126/science.1165243 (2009)
The detection of methane in the atmosphere of Mars five years ago was a surprise, because the gas is not stable under Martian conditions. Michael Mumma at NASA's Goddard Space Flight Center in Greenbelt, Maryland, one of those who made the initial reports, and his colleagues now provide

details about how much methane can be detected on the planet and where.

Using data from two high-resolution spectrometers on Hawaii, they report evidence for periodic, localized plumes of methane in the highland region known as Arabia Terra. By comparing measurements made at different times they estimate that the gas has an atmospheric lifetime of at most a few Earth years. The mechanism of its removal, like the mechanism of its production, is currently unknown.

ENTOMOLOGY

Hammers of the wasps

Biol. J. Linn. Soc. **96,** 82–102 (2009)

A number of parasitic wasp species have independently evolved echolocation techniques to find host insects deep within trees.

Several species of parasitic wasp attack beetle larvae living inside wood, leading researchers to wonder how they find their prey. To this end, Nina Laurenne at the Museum of Natural History in Helsinki and her colleagues have surveyed the hammer-shaped antennal tips that these species whack against the trees. This hammering allows the wasps to locate the regions where they are wont to find their prey.

A phylogenetic analysis conducted by the researchers suggests that these hammers are not a one-off innovation and have appeared and disappeared during the course of evolution, seemingly in response to the needs of wasp species moving into this niche.

NEUROBIOLOGY

Scent slides away

Neuron **61,** 57–79 (2009)

To adjust its behaviour to suit ever-changing environments, an animal's sensory neurons must not only be able to respond to a change — such as a new smell — but also to gauge when the novelty has worn off, and the response is no longer appropriate.

The olfactory neurons of the tiny worm *Caenorhabditis elegans* achieve this by synthesising a key adaptation protein at just the right place and time: in the neuron's sensory cilia when the odour is first encountered.

Noelle L'Etoile from the University of California, Davis, and her colleagues have now shown that this synthesis is increased by RNA-binding proteins known as PUF — which is surprising because PUF proteins suppress protein synthesis during the adaptation of neurons that occurs in the course of development.

CLIMATE

De-fogged

Nature Geosci. 10.1038/ngeo414 (2009)

During the past 30 years, the skies of Europe have become clearer. Robert Vautard of the Laboratory for Climatic and Environmental Science (LSCE) in Gif-sur-Yvette, France, and his colleagues studied records of visibility at weather stations across the continent. They found that low-visibility conditions such as fog, mist and haze have declined by as much as 50% over the period. The pattern of improvement is correlated with local declines in sulphur dioxide emissions, suggesting a role for pollution control.

This brightening of the skies, the researchers say, could have contributed to Europe's reported daytime warming during that time by 10–20%, with a particularly marked effect in eastern Europe.

D. BOHRER, WHITE HOUSE/AP PHOTO

EVOLUTION

Run rabbit run

Proc. Natl Acad. Sci. USA **106,** 952–954 (2009)

Predation by humans drives changes in exploited prey much faster than other evolutionary pressures do.

Previous research has shown that commercial fishers and trophy hunters can mould traits such as average size at reproductive age in wild populations, but no single study has revealed the pace at which these changes generally take place. In a meta-analysis of work on the morphology and life histories of 29 species, including fish, mammals and plants, Chris Darimont of the University of California, Santa Cruz and his colleagues found that changes in human-harvested systems occurred more than 300% faster than in natural systems, and 50% faster than in systems affected by other human influences, such as pollution.

The authors suggest that human predation works so quickly because it is often felt by large proportions of the adults in populations.

JOURNAL CLUB

Paul Knoepfler
University of California, Davis

A cell biologist looks at the risk and promise of a new insight into stem cells and cancer.

I study both stem and tumour cells, and am fascinated by their close relationship. Both exhibit pluripotency — the capacity to develop into any cell type — and the ability to cause cancer. Even some apparently normal stem cells can cause tumours, whereas others, sometimes from the same culture, lack this power. It seems that not all stem cells are created equal — even in the same dish.

A recent paper from Mickie Bhatia's group (T. E. Werbowetski-Ogilvie *et al. Nature Biotechnol.* **27,** 91–97; 2009) is the first to directly address this heterogeneity in human embryonic stem cell (ESC) cultures. The team found that individual human ESC lines contain significant subpopulations that vary in a number of ways, including in tumorigenicity.

Variant human ESC lines were about 20 times more tumorigenic than the cultures they had been derived from and showed small changes in chromosome structure. These could be identified by array-based comparative genomic hybridization (aCGH), but were not detectable by standard karyotyping. Thus for 'normal' stem cells being considered for use in regenerative medicine, karyotyping is not enough. Screening should also include aCGH, and perhaps an analysis of gene-expression patterns.

This previously covert diversity has implications for both tumour biology and medical applications involving stem cells. It may shed light on the 'locked in' self-renewal that is emerging as an important feature of many sorts of tumour and tumour stem cell.

The heterogeneity of human ESC cultures represents an additional hurdle in terms of producing safe stem-cell-based transplants. At the same time, it may offer a valuable bonus: the chance to purify variant human ESC sub-lines that are less tumorigenic.

Discuss this paper at **http://blogs.nature.com/nature/journalclub**

NEWS

Not so sunny after all

Manufacturers in the solar-energy industry are downsizing and scaling back their once-ambitious plans. **Katharine Sanderson** reports.

Lay-offs, stalled projects and grim financial forecasts are affecting the solar-energy industry, which until recently had been growing with little end in sight.

In November 2008, BP Solar announced that it would be closing a photovoltaics factory in Sydney, Australia, so that it could focus its operations in countries with lower manufacturing costs. Since then, GT Solar, of Merrimack, New Hampshire, let go 25 of its 300 workers; Day4Energy in Burnaby, Canada, slashed one-third of its staff, some 95 employees; and even Suntech, one of China's largest solar companies, in Wuxi, Jiangsu province, shed 800 jobs, or 10% of its work force.

One of the most dramatic cutbacks has come at OptiSolar of Hayward, California. The company has been planning to open the largest photovoltaics power plant in the world, a 550-megawatt facility in San Luis Obispo county in California, by 2013. OptiSolar says it still intends to build the plant even though it has been unable to raise money in its latest funding rounds. "We suspended construction in November as it became obvious that access to capital was not available," says spokesman Alan Bernheimer. Nearly half the company's workforce, or 300 employees, will be let go.

Some legislative incentives that could help the industry are on the horizon. Last October, the US Congress extended for another eight years the 30% tax credit for those who invest in renewable-energy projects. But many investment firms aren't making enough money to take advantage of the credit. "This is probably the biggest problem for US industry," says Jenny Chase, a London-based analyst for New Energy Finance. The Solar Energy Industries Association in Washington DC has called for the incentives to be made refundable so that investors can claim them even if they are not making money.

"This was always likely to be the year when supply overtook demand."

The economic stimulus package proposed by the US House of Representatives last week (see page 364) doesn't include such a plan, but it does contain a US$8-billion loan guarantee programme for renewable energy technologies. If passed, this proposal could help OptiSolar, which is applying now for loan guarantees from the government. "We seem to be just what they're talking about," says Bernheimer. Still, even Monique Hanis, director of communications for the industry organization, says that "these are just proposals. There's a lot of work to be done."

Current tax incentives also don't cover solar-cell manufacturers, notes B. J. Stanbery, chief executive of HelioVolt, a company in Austin, Texas, that manufactures copper-indium-gallium-selenide solar cells. HelioVolt, which is just about to bring its first manufacturing plant online, has also made some employees redundant, but won't say how many. Stanbery says that the redundancies were part of the process of shifting from research and development to manufacturing, rather than because of the economic crisis. "Frankly, it had more to do with adjusting the skills mix in the company," he says.

Stanbery predicts that his company will survive as it is young and small — it expects to generate 20 megawatts this year from its production facility in Austin. "I don't think it's going to be a really tough year for us," he

Translational research in Berlin hits a roadblock

A historic biomedical research campus in former East Germany is retrenching after a prolonged political attack. Unproven accusations alleged that €15 million (US$20 million) of state money had flowed illegally into the private Helios hospital in the Berlin suburb of Buch. Clinical researchers there now find themselves struggling to rebuild collaborations with basic scientists.

The dispute concerns a contract forged in 2001 between a national research centre, the Max Delbrück Center for Molecular Medicine, and two hospitals: Helios, which is private, and the Charité, which is Berlin's teaching hospital. The aim was to promote translational research. But the hospitals terminated the contract in December 2008 "because we felt that the scandalizing would never stop unless we did", says Karl Max Einhäupl, head of Charité. Einhäupl says that dividing funds between research and patient care is always complicated, but that he doesn't think any state money was used inappropriately.

Buch developed over the past century as a state-of-the-art biomedical and clinical campus. After reunification, the former West German research system absorbed the basic-research institute but not its clinical facilities, which were eventually privatized.

Keen to improve the parlous state of translational clinical research in Germany — something that East Germany had done well in Buch — the German government helped to facilitate an agreement with the private sector. Clinicians from Charité's medical school were appointed to head clinics at Helios, and their salaries were shared between the hospitals. This gave clinicians with research interests convenient access to the laboratory facilities at Buch.

CHARITÉ

The Buch biomedical campus in northeast Berlin is home to large lab facilities.

HKUST

INTERVIEW: PAUL CHU
Hong Kong inaugurates Institute for Advanced Study.
www.nature.com/news

R. NICKELSBERG/GETTY IMAGES

Solar panels are now being made faster than they can be sold.

says. For others in the solar industry at large, he predicts, "it is going to be brutal".

Solar's rapid growth in recent years has coloured expectations for the field, says Ken Zweibel, director of the Institute for the Analysis of Solar Energy at George Washington University in Washington DC. "Solar has been on a 40–50% growth rate for ten years now," he says. "What we're losing is that growth rate."

This year was always going to be tough for solar, says Chase. In Europe in particular, four years of generous subsidies have encouraged new players to join the fray, leading to a surge of solar modules on the market. "This was always likely to be the year when supply overtook demand," she says. In Spain, the government's incentive scheme was so popular that roughly 3 gigawatts of solar-power capacity were connected to the grid in 2008, and authorities are investigating reports of fraudulent hook-ups. The Spanish government has since capped this year's production subsidies at 500 megawatts.

In Bitterfeld-Wolfen, Germany, solar-cell manufacturer Q-Cells closed its facilities temporarily over the Christmas–New Year period to save money. Q-Cells has also put on hold plans to build a silicon wafer and ingot plant in Malaysia. "We will have to focus on our core business — increasing cell production," says says spokesman Stefan Dietrich. "There's not much money around."

In December, Q-Cells reduced its forecast net income for the 2008 financial year from €215 million (US$286 million) to €185 million. Similarly, the Chinese company LDK Solar, in Xinyu, Jiangxi province, cut its 2008 forecast from a maximum in revenue of $565 million to a maximum of $435 million.

One thing is sure, says Chase: the prices of solar modules will plummet in 2009 because of the oversupply. But low prices could mean that the developing world gets better access to solar power, or attracts investors to companies that had been flagging. "I think the return on solar will go up," she says. It's just that some companies won't make it that long. ■

But Germany has always been uncomfortable with public–private partnerships, particularly within the cash-strapped health system. For more than a year, politicians had been piling pressure on the hospitals, suggesting that public money was seeping into patient care under the guise of research. Two investigations by the consultancy firm PricewaterhouseCoopers failed to disentangle the payments, but concluded that there was "no badly intentioned, deliberate cross-transfer of funds".

Now that the contract has been terminated, clinic heads at Helios have had to decide whether to stay with the hospital and lose their academic status, or move into the city at Charité and lose convenient access to the Buch laboratory facilities.

In the future, the Max Delbrück Center will collaborate with Helios on a project-by-project basis with much paperwork to specify who pays for what.

"The flow of finance will be clearer this way, but [the new bureaucracy] doesn't make our lives particularly pleasant," says Frederick Luft, a nephrologist and hypertension expert who decided to stay with Helios so that he could stay in Buch. "It makes a big difference to be able to walk 10 minutes between the hospital and the lab," he says.

Many of Luft's colleagues also believe that the distance between suburban Buch and the downtown Charité clinics, which can take up to an hour on public transport, makes collaborations less attractive.

"Translational research works best when clinicians and basic scientists meet on a daily basis," says Detlev Ganten, founder of the Max Delbrück Center, former Charité chief and architect of the public–private collaboration contract in 2001. Still, he says, "Buch is not as disconnected as it was, and it will be possible to get things to work again."

Another part of the effort to promote translational research at Buch has also run into trouble. A publicly funded Experimental and Clinical Research Center had been planned to be built on Helios land in 2001, but will now be built a few hundred metres away, next to the Max Delbrück Center, instead. The change will delay its planned 2011 opening by at least a year. Collaborations between the Max Delbrück Center and the Charité, both of which are publicly funded, will now intensify and be organized through the Experimental and Clinical Research Center.

"Translational research works best when clinicians and basic scientists meet on a daily basis."

This week, the Buch campus sees at least one piece of good news. Research minister Annette Schavan is scheduled to open an ambitious new €10-million magnetic resonance imaging facility at the Max Delbrück Center. It includes a 7-tesla imager for use in humans, and a 9.4-tesla imager for animals.

Aside from their more conventional use to look inside the brain, the machines will also be used to develop methods for imaging the beating heart, a challenge very few groups around the world are attempting. The facility is meant as another step in Germany's embrace of translational research in medicine. "Imaging is an important bridge between basic research and medicine," says pharmacologist Walter Rosenthal, scientific director of the Max Delbrück Center. ■

Alison Abbott

Cash boost for US science

Researchers in line for $13-billion windfall.

After getting their first glimpse of the massive financial stimulus bill last week, US researchers are scrambling to work out how to get some of the billions of dollars proposed for science and technology into their laboratories.

On 15 January, the House of Representatives released details of its proposed US$825-billion economic stimulus bill. Along with other spending initiatives and a raft of tax cuts, the blueprint includes new, one-off funding for federal research and development that totals more than $13 billion.

Among the big winners is the National Science Foundation (NSF), which would receive an additional $3 billion — half of its annual budget — of which $2 billion would go directly to research grants. The National Institutes of Health (NIH) would receive $3.5 billion, of which $1.5 billion would be for research at NIH centres over two years; $1.5 billion for building grants at university research facilities; and $500 million for construction on the NIH campus in Bethesda, Maryland. The Department of Energy's Office of Science would receive $2 billion, which includes $400 million to kick-start the Advanced Research Project Agency-Energy, which is meant to fund high-risk research in innovative energy ideas.

As expected, other clean-energy initiatives also reaped billions of dollars. Some $4.5 billion would go to efforts to develop a smart electricity grid; $8 billion would go to loan guarantees for renewable-energy technologies (see page 362); and $2.4 billion would go to projects in carbon capture and sequestration at fossil-fuel-burning plants.

Many research advocates had pushed Congress and the incoming administration of Barack Obama for such investments, and predictably applauded the proposal (see *Nature* **457**, 240–241; 2009).

Robert Berdahl, president of the Association of American Universities in Washington DC, said the bill represents a solid endorsement of the scientific community's argument that investing in research and education provides

House Appropriations chairman David Obey unveiled the winners in the economic stimulus bill.

S. J. FERRELL/CONGRESSIONAL QUARTERLY/GETTY IMAGES

Ebola outbreak has experts rooting for answers

When the Ebola Reston virus was discovered in pigs in the Philippines last year, it marked the virus's first known foray outside primates, and raised fears of a potential threat to human health.

Last week, a joint mission of 22 international health and veterinary experts returned from investigating the outbreak with more questions than answers about the virus's pathology and epidemiology.

The Ebola Reston virus was first discovered, in 1989, in crab-eating macaques imported to the United States from the Philippines. Since then, the virus has killed most infected monkeys, yet had no effect on the 25 people that it infected — unlike three of the four other strains of Ebola, which kill between 25% and 90% of the humans they infect.

Because few people come into close contact with primates in the Philippines, the risk of catching Ebola Reston in this way is relatively low. By contrast, the appearance of the virus in an important livestock species was unexpected and worrying, says Pierre Rollin, an Ebola expert at the US Centers for Disease Control and Prevention (CDC) in Atlanta, Georgia, who was part of the mission to the Philippines. "We never thought that pigs could be infected," he says.

Once inside the pig, it may be possible for the virus to mutate into a version that is deadly to humans, as the avian influenza virus is thought to have done. "And we still don't know what it might do to someone who is immunocompromised by HIV or by drugs," Rollin adds.

But there seems to be little threat to human health from the current form of the virus. It is destroyed by cooking, and there is no evidence of symptoms in pig handlers, who will soon be tested to find out if they have developed antibodies to the virus.

Is an Ebola-virus subtype killing pigs?

REUTERS

The investigation into the Ebola Reston infections began after farmers in the Philippines reported high mortality rates in their pigs in 2008. In September, samples from 28 dead pigs were sent to the Plum Island Animal Disease Center in New York, where researchers found evidence of the porcine reproductive and respiratory syndrome virus, also known as blue-ear pig disease, which has seen many outbreaks in Asia in recent years. But in six of the samples they also found Ebola Reston. This virulent, biosafety-level-4 pathogen requires special laboratory facilities, so the pig samples were rushed to the CDC labs in Atlanta for further analysis.

Despite the presence of other diseases in the samples — including swine fever, and the porcine circovirus type II — Rollin thinks that Ebola Reston is to blame for the pigs' deaths because histological samples showed that the virus had pervaded the spleen, similar to its mode of attack in

HAVE YOUR SAY
Comment on any of our News stories, online.
www.nature.com/news

jobs while laying the foundation for a cleaner, more competitive economy. The trick, he says, will be getting the bill through Congress and then sustaining funding into the future.

"We hope that this gets built into the base and that it is essentially front-loading some of the increases that are planned for 2010 and beyond," says Berdahl. Obama is scheduled to present his proposed budget for fiscal year 2010 in early February.

But some have questioned whether the one-time infusion of cash will matter much to agencies whose budgets have flatlined or been lower than expected in recent years. Elias Zerhouni, former director of the NIH, says the stimulus package does not focus enough on sustaining scientists, concentrating instead on the facilities that house them. The current proposal "is too timid and not strategic enough in addressing the long term", he says. "It's short-term wise but long-term ineffective."

"Researchers out there could actually spend a lot of money wisely and quickly if it were made available."

Democratic congressional leaders, spearheaded by House Speaker Nancy Pelosi of California, developed the bill in consultation with Obama's transition team. Although the legislation is sure to change as it moves through the House and the Senate, the fact that it has the tacit approval of both Pelosi and Obama means that it is likely to survive in some form. Republicans, however, were not consulted.

Most of the stimulus spending would extend over two years, although money for peer-reviewed grants must be awarded within 120 days to ensure it is spent quickly. That could mean that agencies use the money to fund peer-reviewed grants that previously scored highly but were not funded because of a lack of money at the time. At the NSF, individual directorates are likely to be allowed to determine how they will spend the $2-billion windfall. Other chunks have been designated for specific programmes: $400 million, for instance, will go to the major research equipment and facilities programme, which includes large projects that must be approved by the National Science Board. Another $500 million will go to instrumentation, including modernization and retrofitting.

At NASA, the $400 million targeted for the science office includes $250 million to accelerate Earth-sciences projects recommended in a recent prioritization survey by the National Academies. Among other things, that would pay for a climate sensor measuring total solar irradiance to be put back on the next generation of US weather satellites; it had previously been removed to save money.

Bart Gordon (Democrat, Tennessee), chairman of the House Science Committee, called the bill a long-delayed down payment on the American Competitiveness Initiative, which seeks to boost research and education in mathematics, engineering and the physical sciences.

Neal Lane, a professor at Rice University in Houston, Texas, and a former science adviser to President Bill Clinton, says agencies such as the NSF and the Department of Energy will probably be able to absorb and spend the additional funding quickly, in part because they have been planning for increased funding under the competitiveness initiative. The NSF in particular has been faced with a backlog of requests and peer-reviewed proposals for facilities and construction of major research equipment. "Those seem like reasonable places to make a quick and early investment," he says. "The researchers out there could actually spend a lot of money wisely and quickly if it were made available to them." ■

Jeff Tollefson; additional reporting by Meredith Wadman and Rich Monastersky.

monkeys. Further pathology tests are due to begin in spring at the Australian Animal Health Laboratory in Geelong, Victoria.

The infected pigs came from several farms on the island of Luzon, and on 13 January, health officials collected blood and tissue samples from hundreds of apparently healthy pigs there. Although Rollin does not expect to find the virus itself in these samples, the pigs may carry antibodies that should indicate an approximate mortality rate associated with exposure.

Rollin suspects that, as is the case with monkeys, the infections resulted from contact with a reservoir of the virus, rather than spreading from animal to animal. In 2005, outbreaks of human Ebola in Gabon and the Republic of the Congo were traced back to colonies of bats (E. M. Leroy *et al.* *Nature* 438, 575–576; 2005). "It's almost certainly the case [in the Philippines]," says Rollin.

The virus is likely to be spread by bat droppings falling into the pigs' feed, and the threat of infection could be reduced by moving fruit trees, where the bats roost, away from pig farms, or by putting roofs on pig enclosures. "We can't exterminate it, we just have to learn how to avoid it," says Rollin. ■

David Cyranoski

GRAPHIC DETAIL

Prices plummet on carbon market

The price of European Union (EU) allowances for carbon dioxide emissions has reached an all-time low, hit by falling oil and gas prices, and expectations that economic recession will lead to reduced energy demand.

Under the EU's mandatory emission trading system — set up in 2005 and still by far the largest such scheme in the world — power plants and other CO_2-intensive industries can buy emission allowances that allow them to exceed their government-allocated CO_2 caps. In 2008, the equivalent of almost 5 billion tonnes of CO_2 was traded on the global market, an 83% rise on 2007.

At close of trading on 19 January, allowances to emit one extra tonne of CO_2 in 2009 were selling at just €11.65 (US$15.32) on European exchanges that trade carbon, such as the European Energy Exchange (EEX) in Leipzig, Germany (see graph). The price of allowances saw a recent peak of above €30 in July 2008, but falling oil and gas prices have encouraged electricity generators to switch from burning coal to cleaner natural gas, reducing demand for emission allowances.

SOURCES: EEX; TECSON

Given the dire economic outlook, analysts believe that it will be difficult to stop the market's negative trend in the next few months.

The EU's emission trading system is due to be reformed in 2013 to reduce the amount of free allowances handed out to companies (see *Nature* **456,** 847; 2008).

Quirin Schiermeier

SPECIAL REPORT

Which moon to shoot for?

Planetary scientists have a rare chance to pick the destination for their next big mission. But will it be Titan or Europa? **Eric Hand** reports.

In 2005, Jonathan Lunine, a planetary scientist at the University of Arizona in Tucson, was one of the first dozen people to glimpse the surface of Saturn's moon Titan. As the Huygens probe fell, a bizarre landscape emerged: an icy world of hills and channels carved by liquid methane. "And I had two emotions, one right after the other," he says. "The first was elation, astonished elation! And the second one, almost immediately, was sadness. Because I knew that I wouldn't get to see this again in my scientific lifetime."

But Lunine might just be that lucky — if NASA and the European Space Agency (ESA) choose to return to Titan two decades hence. By the end of this month, agency officials plan to pick a destination for a massive mission, costing nearly US$4 billion, to be launched around 2020 for the distant reaches of the Solar System. The battle pits Titan, which recent discoveries have made the cool new kid on the block, against Jupiter's moon Europa, which has long sat atop community wish lists.

> "Are you going to pick the brightly coloured horse, or are you going to ask, 'Is that horse ready to run?'"

The impending decision — a big deal for a science community that typically sees its projects come to fruition just once a generation — has planetary researchers taking sides. "I think it's going to come down to a matter of the science," says Lunine, co-chair of a study pushing for a visit to Titan. His mission, described in glossy brochures and snazzy computer simulations, includes a hot-air balloon straight out of a Jules Verne novel, which would drift in Titan's cold breeze.

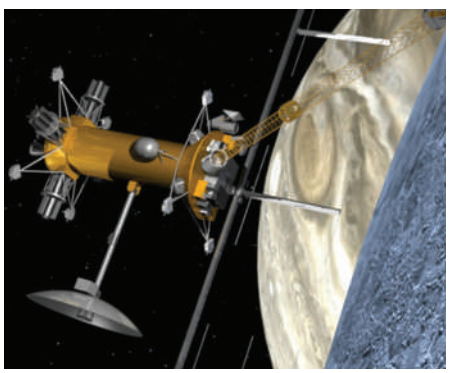
On the other side are the Europa supporters, who argue that a decade of mission-design work should carry the day. They envisage a NASA mission to Europa combined with an ESA mission to Ganymede, another Jovian moon with an intriguing magnetic field. Bob Pappalardo, a planetary scientist at the Jet Propulsion Laboratory (JPL) in Pasadena, California, has worked on five Europa mission studies over a decade; the Titan mission concepts, he says, just recently got going. "Are you going to pick the cool, brightly coloured horse, or are you going to ask, 'Is that horse ready to run?'" he says. "These reviews should be looking at the teeth and gums and seeing which one is ready to go."

Ten years ago, a mission to Europa would have been the brightly coloured horse. In 1995, the Galileo probe began an 8-year tour of Jupiter's system, during which it snapped the first close-ups of Europa's scarred surface. Analysis of a magnetic anomaly soon revealed the moon's most astonishing feature: that eggshell of ice is thought to enclose a warm, salty ocean. Scientists immediately clamoured to return. The JPL began a Europa Orbiter mission study in 1999; Europa was ranked a top priority in an important 2002 community list; even Congress told NASA, in 2005, to begin a Europa mission.

The need to go back is still great, says Pappalardo. Galileo made only 11 passes of Europa and was hampered by an antenna that failed to fully deploy, limiting its data transmissions. An orbiter equipped with new instruments — in particular, an ice-penetrating radar — would end debate over the thickness of the ice shell, and reveal the depth of its fissures and cracks. Astrobiologists want to know how isolated any ocean might be from the surface, where withering radiation would make life tough but could also lead to chemical exchanges that nourish life.

Safe from harm

Radiation would damage not only life but spacecraft electronics and, over the past decade, engineers have developed technology to combat it. The earlier Europa studies funded research in radiation-hardened, or rad-hard, electronics that are now used on military satellites and NASA spacecraft, making the components cheaper. Where rad-hard components still don't exist or are too costly, engineers will nestle the orbiter's 11 proposed instruments behind metallic shields of tantalum and tungsten.



NASA

A trip to Europa would mean building an orbiter...

Still, 20% to 30% of the mission's cost is for radiation protection, according to Europa study leader Karla Clark of the JPL. The Europa orbiter must fit within a NASA cost envelope of $2.9 billion, while the ESA contribution, the Ganymede orbiter, must cost less than €650 million ($860 million). At those prices, a landing element, or even an ice-drilling cryobot, is impossible. But Pappalardo argues that sending a Europa orbiter now could pave the way for a future lander — by scouting for the smooth pavement. "We're ready to go to Europa now," he says, "and we'll be ready to do Titan next."

Jupiter, then Saturn, then Europa, then Titan — this notion of ordered 'turns' doesn't fly for Ralph Lorenz, a planetary scientist at the Johns Hopkins University Applied Physics Laboratory in Laurel, Maryland. By that logic, he says, it's Neptune's turn; it's Venus's turn. "You've got to choose the mission that's giving you the best science value for the money," he says. "And Titan is the low-hanging fruit." With its atmosphere, its hydrological cycle and even its potential for cryovolcanism, Titan would mobilize a much wider scientific community, he says.

NASA would send a spacecraft into polar orbit for four years, where it would spend more time at Titan in its first three days than the Cassini mission to Saturn will in its entire mission. To keep costs down, the orbiter would carry six instruments — about half as many as Cassini. But they would be fine-tuned for Titan, says Lunine. A mass spectrometer, for instance, would be sensitive to molecular chains with hundreds of carbons — orders of magnitude more sensitive than Cassini's spectrometer, which is limited to molecules with seven or eight carbons.

After arrival, ESA would showcase its contributions. The first — a dropped lander akin to Cassini's Huygens probe — would splash down into one of two giant methane-ethane lakes near Titan's north pole. It would float for a few hours in the complete darkness of winter and suck up samples through straws, sifting for the molecular evidence of organized organic chemistry that, even if pre-biotic, so excites astrobiologists. "You go to the lake," says Lunine, "because the lake is nature's great

M. HEGNER

NANOTECHNOLOGY
Tiny springboards detect viruses in fluids.
www.nature.com/news

NASA JPL/CORBY WASTE

...but a mission to Titan could feature a hot-air balloon.

organic solvent medium for delivering these components to you."

But the *pièce de résistance* is the 11-metre-wide hot-air balloon, which, drifting along 10 kilometres high in gentle winds near the equator, would image the surface through the methane mist. Heated by lumps of radioactive elements, the balloon would circle the moon at least once in a lifetime of a minimum of six months, says Athena Coustenis of the Paris Observatory, the lead European study scientist for the Titan mission. The heat shield used during the balloon's initial descent would not be wasted: there is a proposal to have it stick gently in the soil as a 'geosaucer', with a seismometer and magnetometer to measure tidal flexure and cryovolcanic rumblings. The balloon is likely to be built in France where, says Coustenis, "balloons are a big deal".

In the end, the decision between Titan and Europa could come down not just to scientific promise versus technological readiness, but also to agency politics. The decision-makers, Edward Weiler and David Southwood, science chiefs for NASA and ESA respectively, haven't been shy in expressing personal opinions. "I've seen the data coming out of Cassini and it knocks your socks off," says Weiler. "I just have a preference for Titan." And Southwood says, "I have an emotional preference for Saturn and Titan. But I probably have a scientific longing for Jupiter."

Early decision

After reviewing the mission studies and reports from the competing teams, Weiler and Southwood will hold a teleconference by the end of the month in the hope of making a decision before an ESA advisory committee meeting on 3 February. Weiler says technological readiness will probably be a deciding factor. If both missions are deemed technologically sound, Weiler says he may ask a panel convened by the US National Academies for advice on the most scientifically worthy moon.

"Balloons are a big deal in France."

At ESA, the chosen mission will still have to compete against two other missions vying for one €650-million slot within the agency's Cosmic Vision programme, a competition not due for final selection until 2011. But Southwood says that ESA needs to decide now whether it wants to pursue Titan, or Ganymede during a NASA mission to Europa. That's because the technologies and approaches needed for either moon are so disparate. Titan would mean developing a lander and balloon linked tightly to the US orbiter. Ganymede would mean developing an orbiter to be launched separately from the US Europa probe. And should Europe decide, come 2011, that some entirely different mission should get the Cosmic Vision slot, then NASA would be left to either pick up some of the European ideas for its own Titan or Europa mission, or to scale back its own ambitions.

Although proposal scientists are nervously anticipating the decision, they will have a lot longer to wait for the mission itself. A probe launched for Jupiter in 2020 wouldn't arrive in Europa orbit until 2028, and a balloon bound for Titan wouldn't deploy until 2030. Lunine would be 71 years old — not necessarily beyond the workplace and, he hopes, spry enough for a little ballooning. ■

See Editorial, page 358.

A fly by any other name

Drosophila experts argue over reclassification proposal.

The Byzantine world of species taxonomy is facing a new test: a proposal that would involve renaming *Drosophila* flies, arguably the leading model genetic organism.

The idea is to bring a new taxonomic order to the more than 2,000 species now grouped in the genus *Drosophila*. An application pending before the International Commission on Zoological Nomenclature (ICZN) would designate *Drosophila melanogaster* — the primary species used for genetics studies — as the type species for the genus. Currently, the type species is *D. funebris*, which was described in 1787 by Johann Fabricius.

Kim van der Linde, a postdoctoral researcher at Florida State University in Tallahassee, and colleagues filed the application in December 2007. Recent genetic work has shown that *D. melanogaster* could be classified as *Sophophora* and could therefore be renamed *Sophophora melanogaster*. Van der Linde says her proposal is designed to preserve the name *D. melanogaster*.

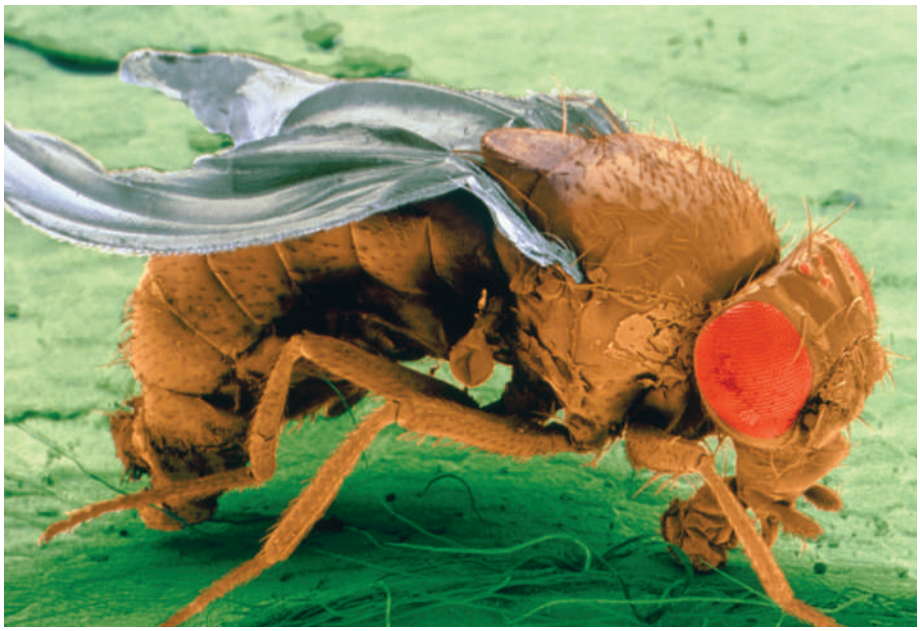
Launching the idea to update the nomenclature has let the fly-naming conundrum out of the bottle, touching off what ICZN staff call an unprecedented debate. Some worry that teaching, citation searches, publications and databases such as FlyBase would face a logistical nightmare if *D. melanogaster* changed its name.

"If this were some obscure beetle, you could rename it Godzilla and it wouldn't make much difference," says Therese Markow, a geneticist at the University of California at San Diego who directs the *Drosophila* Species Stock Center. "But this is the premier genetic model system." In November 2008, Markow convened a workshop in San Diego where many of the world's leading *Drosophila* authorities wrestled with van der Linde's proposal, without resolve.

"If this were some obscure beetle, you could rename it Godzilla and it wouldn't make much difference."

Markow supports the status quo. So does Thomas Kaufman, a *Drosophila* geneticist at Indiana University in Bloomington and co-leader of FlyBase. "We should leave the naming alone — everything is working fine," he says. "We can't even get agreement on names for the mutants, let alone the whole genus." A number of formal comments on the ICZN petition were strongly opposed to it.

Still, some taxonomists say it's time for reordering, especially given the recent sequencing of many *Drosophila* species, such as the 2007 publication of a dozen genomes that show how closely related species are to each other on a genetic level[1,2]. Van der Linde, a community ecologist trained in the Netherlands, got started on her proposal after doing some fieldwork in the Philippines. She joined up with several colleagues, including the *Drosophila* taxonomist Masanori Toda of Hokkaido University in Japan, and naming authority Gerhard Bächli of the Zoological Museum in Zurich, Switzerland.



EYE OF SCIENCE/SPL

Name game: Should the *Drosophila melanogaster* fly be the type species for the genus?

The group developed a petition (see http://tinyurl.com/999mep) for the ICZN, and wrote a manuscript prescribing their interpretation. In renaming, a scientist can seek the approval of the commission, which then publishes the decision of its panel of about 25 authorities. Or a scientist can publish an article proposing the new order, leaving it to scientific communities to adjust names. "People can follow or reject us," says van der Linde.

A decision from the commission could take another year or more. If it rejects the proposal, taxonomists could take it on themselves to split the genera in descriptions in future publications. Under this scenario, *D. melanogaster* could become *S. melanogaster*, and 1,000 other species could be renamed too.

Earlier this month, van der Linde said she received commission feedback suggesting the petition may be set aside. That would leave the community to split the genus at will. This approach is supported by the international editorial team of the BioSystematic Database of World Diptera, a US-based fly name resource. In a comment on van der Linde's petition, they note that today's computer search engines "will have no trouble finding information" if *D. melanogaster* were to become *S. melanogaster*.

Patrick O'Grady, a geneticist at the University of California at Berkeley who does not support the van der Linde proposal, disagrees. "There would be chaos in the literature if *D. melanogaster* changes to *S. melanogaster*," he says. "Genetic work could be lost. It would be hard to find things." As an example, he points to the 2004 renaming, through publication[3], of the mosquito that spreads yellow fever and dengue: *Aedes aegypti* to *Stegomyia aegypti*. Some researchers cite the new name; others ignore it and use the old name[4]. ■

Rex Dalton

1. Stark, A. *et al. Nature* **450,** 219–232 (2007).
2. Markow, T. A. & O'Grady, P. M. *Genetics* **177,** 1269–1276 (2007).
3. Reinert, J. F., Harbach, R. E. & Kitching, I. J. *Zool. J. Linn. Soc.* **142,** 289–368 (2004).
4. Polaszek, A. *Trends Parasitol.* **22,** 8–9 (2006).

No bull: genes for better milk

On 13 January, the US Department of Agriculture (USDA) launched a service that allows dairy-cattle breeders to double their chances of selecting the best bulls to sire milk-producing cows.

"This is the future of animal breeding," says Juergen Richt, a veterinary surgeon at Kansas State University in Manhattan.

For a decade, breeders who want to locate the best bull have the animals' semen tested for its DNA, looking for traits linked to milk quality and production. About a year ago, the leading artificial-insemination organizations in the United States and Canada funded a US$1-million research project directed by Curtis Van Tassell, a geneticist at the USDA's Bovine Functional Genomics Laboratory in Beltsville, Maryland. Working with Illumina Inc. of San Diego, California, Van Tassell's team created a microarray chip containing 54,000 genetic markers called single nucleotide polymorphisms, or SNPs, that involve at least a dozen traits, including those known to affect milk quality and production.

J. GREEN/REUTERS/CORBIS

DNA testing can predict the best milk producers.

Using high-throughput analysis, the researchers could then compare the DNA from a young dairy bull against the chip SNPs, telling breeders which bull would be likely to sire calves that were good milk producers. The test costs about $225, and can be done when a bull is born, thus avoiding the $25,000–50,000 cost of raising a bull for five years to see if it sires good milk-producing offspring. "The best bulls become elite breeders," says Van Tassell, "The others become hamburger."

Previously, DNA tests allowed a typical breeder to select the best bull some 35% of the time, says geneticist Ole Meland, vice-president of Accelerated Genetics in Baraboo, Wisconsin. The new technique identifies the best bull 70% of the time.

The US initiative is the first such nationwide programme. Companies in New Zealand and the Netherlands have set up private services for cattle breeders; and, following the USDA's lead, similar systems are being built by researchers at Aarhus University in Denmark, and in France and Australia. ■

Rex Dalton

Plan to grow more cannabis for research turned down

US researchers do not need a second federally approved facility for providing research-grade cannabis, the Drug Enforcement Administration (DEA) has ruled. The decision ends an eight-year bid by Lyle Craker at the University of Massachusetts in Amherst, to grow cannabis for medical research, which was supported by the Multidisciplinary Association for Psychedelic Studies (MAPS). A single laboratory at the University of Mississippi supplies cannabis under a contract with the National Institute of Drug Addiction.

Craker, MAPS and others have argued that the current supply of cannabis is of inconsistent quality, difficult to obtain, and that efforts to widen that supply are being stalled for political reasons (see *Nature* 430, 492; 2004). In 2007, a DEA judge recommended that Craker's application to set up a second facility should be granted; but the DEA's final ruling, made on 7 January, said the current supply was adequate.

FDA to regulate the use of transgenic animals

The US Food and Drug Administration (FDA) has adopted guidelines for regulating genetically engineered animals.

Released on 15 January, the rules have been more than a decade in the making. They follow a set of draft guidelines, released in September 2008, which drew fire for effectively treating genetically engineered animals as drugs (see *Nature* 456, 2; 2008).

In response to other concerns about the transparency of the approval process, the new guidelines note that the agency intends to hold public advisory committee meetings before approving any genetically engineered animals.

Earlier this month, an FDA advisory committee deemed an anti-clotting drug called ATryn, produced in the milk of genetically engineered goats, to be safe — an important step towards the eventual approval of the drug for sale in US markets.

For a longer version of this story, see http://tinyurl.com/8kv5dr.

GTC

Goats: drug factories of the future?

Europe set to clamp down on pesticide use

WWW.JUPITERIMAGES.COM

After three years of wrangling, the European Union (EU) is expected next month to finally approve controversial legislation to regulate pesticide use. Early drafts of the law drew protest from manufacturers, farmers and scientists, who claimed that a drastic reduction in the number of available pesticides would lower crop yields and raise food prices.

The compromise ruling backed last week by the European Parliament will result in just 23 of the roughly 500 marketed pesticides being banned, according to the Swedish Chemicals Agency.

The directive also promotes the use of non-chemical pest-control methods, bans aerial crop spraying without authorization and curbs pesticide use in areas such as parks and playgrounds. Once approved, the ruling must be incorporated into the national laws of all member states by 2011.

For a longer version of this story, see: http://tinyurl.com/7d6zvt.

Polio eradication battle is bolstered by $630 million

A final push to overcome the few remaining pockets of polio infection in the world is being supported by grants totalling US$630 million from the Bill & Melinda Gates Foundation, the humanitarian group Rotary International, and the UK and German governments.

The funding, announced on 21 January, will support the World Health Organization's Global Polio Eradication Initiative, which involves vaccinating billions of children. Since its launch in 1988, the effort has cut the annual polio toll from 350,000 cases in 125 countries to just 1,600 cases last year. If the crippling disease could be eradicated, it would become the second disease, after smallpox, to be officially wiped out.

Reservoirs of polio now remain in four countries where the disease is endemic: Nigeria, where polio vaccines were opposed following false rumours that they carried HIV and caused female infertility; India, where vaccine effectiveness has been hampered by poor sanitation and high population density; and conflict zones in Afghanistan and Pakistan.

For a longer version of this story, see http://tinyurl.com/7e5s4u.

Democrats hasten action on climate legislation

Democrats in the US House of Representatives have announced plans to draw up climate legislation this spring, arguing that business leaders need regulatory certainty to drive a green economic recovery.

Incoming Energy and Commerce Chairman Henry Waxman, who held his first climate hearing on 15 January, said that the committee would deliver the legislation by the end of May, a decision backed by House Speaker Nancy Pelosi. Although President Obama favours a domestic cap-and-trade programme, many energy and climate analysts expected that the legislation would be delayed as lawmakers focus on the economy and a less controversial bill to advance clean energy.

The US Climate Action Partnership, a coalition of major industrial firms and environmental organizations, bolstered Waxman's case by calling for legislation that would curb carbon dioxide emissions by 42% by 2030 and 80% by mid-century compared with 2005 levels.

Billion-dollar neutron facility gets thumbs up

The US Department of Energy has approved a US$1-billion upgrade to the Spallation Neutron Source (SNS) at Oak Ridge National Laboratory in Tennessee.

The SNS generates neutrons by firing a high-energy beam of protons at a mercury target. The neutrons are used to probe the structures of materials ranging from proteins to superconducting ceramics. The existing SNS facility, completed in 2006, currently feeds neutrons to 10 instruments used by about 1,000 scientists and engineers last year. The addition of a second target will allow for up to 24 more instruments, fed by long pulses of slower-moving cold neutrons, which can be used to study larger molecules such as polymers.

Conceptual design work on the upgrade can now begin, with construction expected to finish no earlier than 2020.

Beijing's windy bet

After spurning wind power, China has swung around and embraced this clean energy. But the nation's love affair with wind may be spinning out of control, finds **David Cyranoski**.

As he drives along a two-lane highway skirting the Gobi Desert, Hubert Beaumont sees nothing but wind. The road in this part of China's Gansu province runs completely flat — a monotonous route past grey rocks, a few shrubs and some 100 sleek turbines towering over the empty terrain. "I can't think of anything else you'd want to do here except wind energy," says Beaumont, a Beijing-based wind specialist at Ecofys, a green-energy consulting firm.

Indeed, the region is pegged to become the core of a 'mega wind-power base', a massive collection of wind farms with 5 gigawatts of capacity by 2010. Even considering that wind turbines generally produce only about a quarter of their advertised capacity, these would still generate enough juice to satisfy about a million energy-guzzling US homes. It will be the biggest wind-power development in the world.

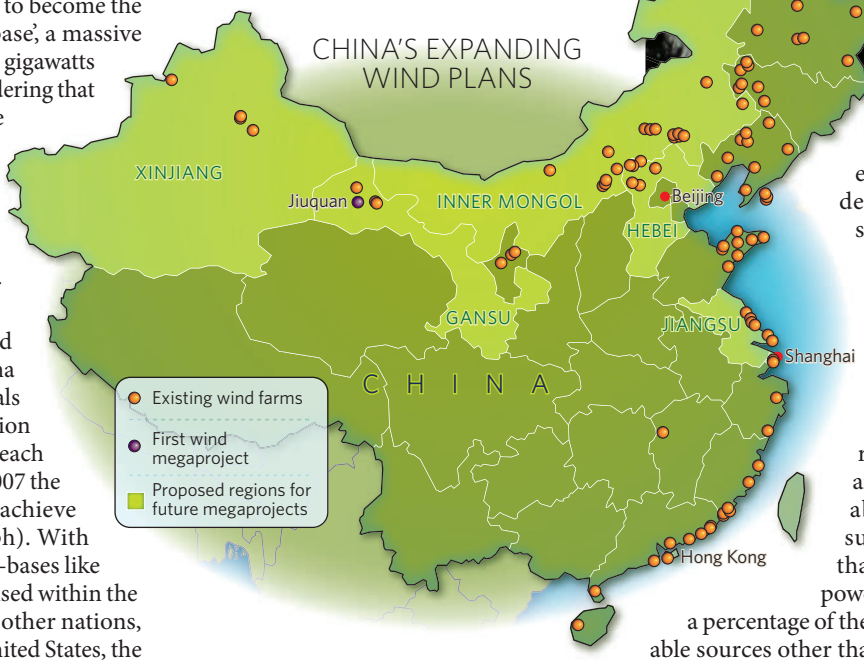
For a country with such a bad environmental reputation, China is fast amassing green credentials in wind. Wind-energy generation capacity has nearly doubled in each of the past three years, and in 2007 the country surpassed its goal to achieve 5 gigawatts by 2010 (see graph). With plans to build four more mega-bases like the Gansu project, China is poised within the next decade to blow past four other nations, including Germany and the United States, the current number one and two in wind-energy capacity.

China is also manufacturing and, increasingly, designing turbines. Domestic turbine manufacturers numbered just four in 2004. Now there are some 70. Within three years, says Haiyan Qin, secretary-general of the Chinese Wind Energy Association in Beijing, China will manufacture more turbines than other country.

But experts wonder whether the superlatives applied to Chinese wind farms will also include 'least reliable' and 'most inefficient'.

Wind-turbine manufacturers and wind-farm developers everywhere have faced teething problems, but China has perhaps faced more difficulties than most. Its wind farms are much less efficient than those in other leading countries, manufacturing defects have plagued Chinese equipment and the nation's electrical grid cannot carry all the wind power the country is generating today, let alone the huge amounts planned for the next few years. Some critics, including several from international turbine companies, blame a lack of planning and poor Chinese manufacturing.

Chinese engineers, entrepreneurs and government officials are working to improve the situation but they have a long way to go. As Beaumont passes the Gansu farm, he sees both the promise and the perils of China's surge in wind power. A strong breeze blows across the desert but half of the turbines are standing still.



Storm chasers

Although China is currently chasing wind power, its leaders had little affection for it just a few years ago. Over the 1980s and 1990s, efforts by the World Bank, the Asian Development Bank and others to get energy from wind in China fizzled out. A goal to have 1 gigawatt of capacity by 2000 passed without notice, or achievement.

"Before 2004, leaders thought wind was too small. They didn't think it was real energy," says Qin. But the sizzling economy boosted demand, energy prices soared and, in 2004, 24 of China's 27 provinces were hit by blackouts. At around the same time, environmental pressures set in ahead of the 2008 Olympics. The government responded in February 2005 with a Renewable Energy Law and subsequent guidelines that called for all major power companies to create a percentage of their energy from renewable sources other than hydropower: 3% by 2010 and 8% by 2020. With biomass resources too sparse and photovoltaics too expensive (although China is the biggest producer in the world), most of that renewable energy will be achieved through wind.

Beginning in 2004, farms started to open up in the windy northern and eastern perimeter of the country (see map). Total installed capacity climbed steeply: 1.3 gigawatts in 2005, 2.6 gigawatts in 2006, 5.9 gigawatts in 2007 and, according to early estimates, 10.6 gigawatts in 2008.

But there is a hitch — China's wind farms are underperforming. All power installations have a 'capacity factor', which is calculated by dividing the energy actually produced by what the installations could maximally generate. According to data from the Beijing branch of

R. PYLE/CORBIS

To meet its renewable-energy targets, China will need to get 5% of its power from wind farms by 2020.

London's New Energy Finance, a consultancy firm that advises investors on developments in renewable energy, on-shore turbines in other leading wind power countries have capacity factors of around 30%. China's is just 23%.

"China's numbers are not good," says the firm's Justin Wu. "It might not seem significant, but a few percentage points could make the difference between a farm that is economically viable and one that is not."

Wind experts blame several factors, starting with the turbines themselves. When wind-farm developers began gearing up around 2004, they imported turbines from established overseas manufacturers. Recently they have relied more on domestic makers. In 2005, Beijing added a requirement that 70% of turbine parts be made by domestic manufacturers. Major international turbine makers have established manufacturing plants in China, but they are losing out to local firms. According to data collected by the China Wind Energy Association, 2008 will be the first time that the installed capacity of Chinese-made turbines will exceed that of foreign ones. The 3.8 gigawatts already assigned to developers for the Gansu project, for example, does not include a single turbine from a foreign maker.

That's great for China's wind-turbine latecomers, but is it good for the wind farms? According to Wu, Chinese wind farms using foreign models have a 5% higher overall capacity factor than those using domestic turbines. Domestic turbines are especially unproductive when first set up, says Wu.

Because the technology is newer and less tested, Chinese turbines are also more likely to be shut down for maintenance, according to anecdotal evidence. A turbine's average down time is a closely guarded trade secret. But Xiliang Zhang, director of Tsinghua University's Institute of Energy, Environment and Economy in Beijing, says he hears reports that domestic turbines are standing still while foreign models nearby are humming away.

Domestic models are 15–20% cheaper, and the quality has been rising quickly as they have either used proven foreign technology or have hired foreign engineers to help with designs. But as the lifetime of a turbine averages 20 years, the more-productive foreign models would be better buys overall, says Wu. (Representatives of two major Chinese turbine manufacturers did not return phone calls or e-mails requesting comments.)

So why do wind-farm developers in China mostly choose domestic turbines? It is a controversial subject. Despite the mandate to source 70% of parts from domestic manufacturers, turbines used in major national projects, including the recently started wind megabases, are supposed to be purchased through open bidding. But the bidding process is a tricky business and many in the industry assume an opaque policy of favouritism, especially considering that most of the wind-farm developers and domestic turbine-makers are state-owned enterprises. "The Chinese government cannot be taken as a model of transparency," says Paulo Fernando Soares, chief executive of Suzlon Energy in Beijing, a subsidiary of a major turbine manufacturer in India.

"The Chinese government cannot be taken as a model of transparency."
— Paulo Fernando Soares

Spin tactics

Turbine quality is one problem; finding the best turbine for the available wind is another. Every turbine has a 'load envelope' that defines roughly what wind speed, turbulence, wind shear and other conditions it will function best in. Some turbines work better for wind that comes in short bursts whereas others work better with long consistent spells of low wind. Poorly chosen turbines will be more likely to break down. A turbine with a wide blade can catch low winds but could be destroyed by typhoons. If turbines are poorly chosen for a site, their efficiency will plummet. "Decisions made in the first year will affect the 20-year life of the turbine," says Sebastian Meyer, who researches wind-energy resources in China for Ecofys. "Serious companies won't put turbines in sites for which they are not suitable," says Soares. But until now, he says, Chinese developers have too often relied on inadequate wind data.

China is now trying to address this problem. By June, it will finish installing 400 masts to measure wind-energy resources in various regions throughout China. The masts will stand 70 metres, 100 metres or 120 metres tall, to match the height of the turbines. China plans to spend more than 200 million renminbi (US$30 million) on the four-year project, which started in 2007. "This project is the key to make the national plan of wind-power development work," says Zhenbin Yang, deputy director of the wind-resource laboratory at the Chinese Academy of Meteorological Sciences in Beijing. Yang will analyse the data provided by the masts.

The plan has its sceptics. Because wind can vary greatly over short distances, even this survey of broad areas of wind flow will be inadequate and 'micrositing' data will still be necessary from the specific regions where the turbines are planned, says Soares.

More scepticism is aimed at a proposed secondary use of the data — to create new standards for turbine design specifically based on Chinese weather. "Wind turbines are not like refrigerators sitting in a house," says Zhang, "China is very different from the United States and Europe." But although it is still not clear what form the Chinese standards might take, some question whether China really needs its own standards. "It's no colder in Mongolia than

Gridlocked: Problems with the electricity grid mean that wind turbines are often at a standstill.

RYAN PYLE/CORBIS

it is in Minnesota," says Soares.

Given China's plans to make five huge megabases, the number of turbines in its future is staggering. By 2015, the Gansu project alone will boast 10,000 turbines with a combined capacity of 12 gigawatts. China's four other megabases — in Xinjiang, Inner Mongolia, Hebei and the Shanghai–Jiangsu region — will total 60 gigawatts.

But to make use of all that energy, China's wind hopefuls must tackle an even more intractable problem — the electricity grid. China's energy supply has been hamstrung by a fragmented and underdeveloped grid system that makes it difficult to get energy from coal-rich rural areas to the cities on the east coast. Wind has the same problem: it is produced mostly in sparsely populated regions that cannot use all that energy. The problem is compounded by the unreliability of a power source that is based on the weather. "Grid companies don't like wind. It changes too fast," says Qin.

Steve Sawyer, secretary-general of the Brussels-based Global Wind Energy Council, says that other countries, particularly the United States, have struggled with their grids. This month, in fact, Democrats introduced a bill in the House of Representatives that allocates US $11 billion to modernize the grid.

> **"Grid companies don't like wind. It changes too fast."**
> — Haiyan Qin

Hurry up and wait

China's rapid expansion has caused delays down the line. Turbines often have to sit idle — on average for four months — before they get hooked up to the grid. The backlog is huge. Of the 5.7 gigawatts of turbine capacity installed by the end of 2007, only 4 gigawatts was plugged into the grid.

Getting connected is just the first hurdle. Many grids are simply too puny to carry all the electricity being made. At peak production times, turbines often have to shut down so as not to overload the electrical networks. Newer turbines can alter the angle of their blades to miss the wind, and will slow to a halt. The only loss is the energy. For older turbines, the operator has to slam on the brakes. Turbines in China wear through their brakes at remarkable speeds, says Meyer. Elsewhere, the brakes are usually used "once a month or less, but in China they're using it every four days", he says.

The delays and losses could drive people away from wind power, says Qin. "If there is not a serious plan for the grid soon, we will not be able to develop wind energy beyond the next few years," he warns.

So why hasn't the Chinese government's legendary ability to get things done kicked in. According to a December 2008 report by New Energy Finance, China's National Development and Reform Commission — the body that oversees national economic and social development — is concerned about the "lack of supervision in China's rapid wind-power growth", especially when it comes to the grid.

But the government is afraid to do anything that would raise prices, says Zhang. The problem is low demand for the expensive energy. "The solution is to force places, like Beijing, to buy it," says Zhang. But the government hasn't brought itself to do that. It's easier to let them stick with coal.

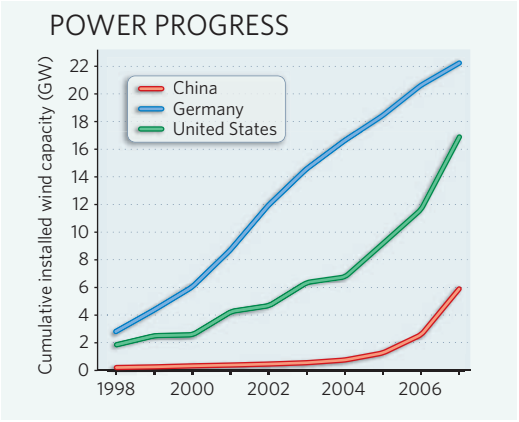
Help might be on the way. The megabases have an economy of scale that will make them more attractive targets for grid developers. And in November last year, the country's biggest power supplier, the State Grid Corporation of China, which provides some 88% of the nation's power, announced plans to more than double its investment in grid infrastructure for 2009 and 2010 from 550 billion renminbi to 1.16 trillion renminbi. Regional projects will also help. In November this year, Gansu will put 20 billion renminbi into its grid to support its megabase, and Inner Mongolia, also set to get a megabase, plans to add 30 billion renminbi to its grid operations by 2010. These will feed into a scheduled upgrade from 110 kilovolts to 750 kilovolts for a transmission line running from Xinjiang in the northwest, through Gansu, to the eastern metropolises. "It gives a feeling of optimism," says Beaumont.

Of course, hooking up wind turbines might not be the main thing on grid developers' minds. Wind energy is currently only a drop in the bucket in China's 793-gigawatt energy supply. In 2008, wind accounted for just 0.3% of China's total electricity production, most of which was powered by coal.

But if the five megabases go as planned, their combined capacity would provide more electricity than the controversial Three Gorges dam, even considering the dam's much higher capacity factor. Meeting government targets for renewable energy for 2020 will, according to most estimates, require a prodigious 100 gigawatts of wind energy — about 5% of China's total energy supply. Some estimates note a potential of 500–600 gigawatts by 2040 to 2050.

For now, the Chinese government has thrown its considerable weight into exploiting this resource. "It is its flagship renewable-energy industry, so the government is going to support it," says Wu. But unless some dramatic changes are made soon, China's plans for wind energy might get blown far off course. ■

David Cyranoski is *Nature*'s Asia-Pacific correspondent.



See Editorial, page 357.

Crisis communication

Messages appear on Internet-based social networks within minutes of disasters occurring. **Lea Winerman** investigates how to harness this trend to create official community-response grids.

The messages began to fly almost as soon as the bullets stopped. Starting at 7:00 on the morning of 16 April 2007, an undergraduate named Seung-Hui Cho had carried a pair of semi-automatic pistols through the campus of the Virginia Polytechnic Institute and State University in Blacksburg — better known as Virginia Tech — gunning down dozens of students and professors as he went. By 9:51 it was over: Cho had turned one of his guns on himself. But the survivors, lacking any official word from the university other than the total death toll, were still in the dark about which of their friends had lived or died. So they turned to the best information source they had: the Internet — notably, the social website Facebook.

Posts appeared in quick succession, indicating the names of suspected casualties. Those here have been edited for privacy reasons. "CH, as reported by a sorority sister," read a post on one Facebook page. "I just finished speaking with his girlfriend, and it appears JH is a fatality as well," read another post. As the information accumulated, the participants spontaneously began to develop their own norms to ensure accuracy. Anonymous or vague posters were asked for clarification. People had to identify themselves when they put forward the name of a shooting victim, for example, and explain where they had got the information.

> **"We saw that the information posted on websites is self-correcting."**
> — Leysia Palen

By the time the university released the names one day later, it was old news to the online community: they had identified all 32 of the deceased already.

The Virginia Tech story is hardly unique. "When people are under threat, perceived or actual, they go into this intensified information-seeking period," explains Leysia Palen, a computer scientist at the University of Colorado, Boulder. And these days, they are increasingly doing so through social networking sites.

But social-network users often end up bypassing the authorities — a tendency that has left officials scrambling to use this information and integrate it into traditional responses.

"Emergency managers have this desire to control the flow of information," says Jeannette Sutton, a sociologist who also studies disaster communication at the University of Colorado. "But you can't control it. The best we'll be able to do is figure out how to harness it."

Palen calls her research 'crisis informatics', and it has taken her to several disaster hotspots. In the Virginia Tech case, for example, she and her colleagues began monitoring websites within hours of the shooting, then travelled to Blacksburg five days later to interview people. Back in Colorado, they created a detailed timeline of official communications (university e-mails, press conferences) and unofficial communications (Facebook messages, Flickr photos) that stretched across a giant white-paper diagram on their office wall.

This allowed them to track the emergence of those informal norms. "One of the concerns from the emergency-management point of view is 'how can we know the information [posted on websites] is accurate?'" Palen says. "We saw that here it was self-correcting."

Spreading like wildfire

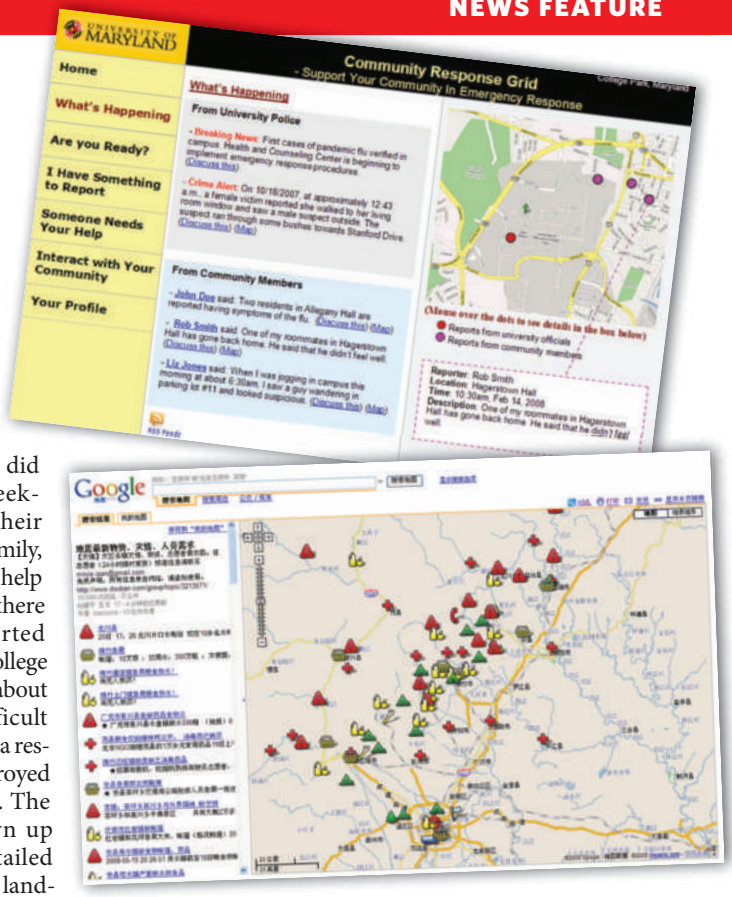
In October and November 2007, Palen and her colleagues examined the online response to another crisis. During that time, more than 20 wildfires raged from Santa Barbara to San Diego counties in southern California, eventually burning 200,000 hectares, destroying about 1,500 homes and forcing many more households to evacuate.

Palen and her colleagues monitored local news websites and online forums including Craigslist, Facebook, Twitter and Flickr. By the tenth day of the fires, they also began to survey and interview area residents.

"National news websites were completely worthless as they ignored everything except the comparatively minor Malibu fire that burned near some celebrity homes," one evacuee wrote.

And official government communications, although sometimes useful, couldn't be relied

FROM LEFT, CLOCKWISE: D. MCNEW/GETTY IMAGES; O. BALLTY/AP; R. L. WOLLENBERG/UPI/NEWSCOM



B. SHNEIDERMAN

Y. QU, P. WU AND X. WANG

Community response grids (top) integrate official responses with information from the public, such as this Google map created by Chinese citizens during the Sichuan quake in 2008.

on either. "The county so-called emergency site was always crashed," another wrote.

Instead, many people turned to websites run by local media, such as the National Public Radio station KPBS, based in San Diego, or by individuals. On these sites, the updates came from any local resident with an Internet connection and information to share. Some hosted Google maps on which users could overlay information such as the location of the fires. Others hosted discussion boards on which people who hadn't evacuated, or who had made it back to their homes, could share what they were seeing.

Worldwide webs

Online information-sharing during crises isn't limited to the United States, and other media agencies are already using the concept. In early January, for example, Al Jazeera, a satellite television network headquartered in Doha, Qatar, launched an experimental website that aggregates text messages, mobile-phone reports and Twitter feeds about the conflict in the Gaza Strip on a Microsoft Virtual Earth map (http://labs.aljazeera.net/warongaza).

And Yan Qu, Philip Wu and Xiaoqing Wang, researchers at the University of Maryland in College Park, observed information-sharing trends after the 12 May 2008 earthquake in China's Sichuan province.

"The earthquake happened at 14:28. Within one minute there was a message posted on a Tianya forum," says Wu. Tianya is one of China's most popular websites — a bulletin-board system with more than 20 million registered users. Within 10 minutes, 56 discussion threads reported feeling the earthquake in 22 cities throughout the region.

The researchers read and classified the thousands of threads that appeared on the site in the days following the quake.

They found that Chinese citizens used the site for many of the same purposes as Californians did during the wildfires: seeking information about their homes, hometowns and family, and coordinating action to help victims of the crisis. And there were some widely reported success stories. In one, a college student saw news reports about the military having a difficult time finding a spot to land a rescue helicopter near a destroyed village in the mountains. The woman, who had grown up in the area, posted a detailed description of a potential landing spot in an online forum, and begged users to forward it to authorities. The post eventually found its way to the military, who landed a helicopter in the spot she described.

Unfortunately, says Wu, that kind of story is rare. In general, professional emergency responders are only vaguely aware of how citizens use social media during disasters. "Even in the United States, professional emergency responders are just beginning to think about this. In China I don't think there's any kind of explicit effort," he says.

It's the lack of official involvement that Ben Shneiderman, a computer-science professor at the University of Maryland and his wife Jennifer Preece, dean of the university's College of Information Studies, are trying to change. In 2007, they published a one-page article in *Science* called '911.gov' (B. Shneiderman and J. Preece *Science* **315**, 944; 2007). Taking their title from the emergency telephone number for much of North America, they envisioned a web-based 'community response grid' that would combine the power of social networking sites with official government emergency-response systems.

An emergency telephone system works terrifically under normal circumstances, says Shneiderman. "But when you get a Hurricane Katrina or a 9/11, [the call centres] become overwhelmed." Potentially, at least, a website can handle such sudden surges more gracefully by tapping more servers as needed.

As Shneiderman sees it, people would report incidents via the Internet or by sending text messages from their mobile phone rather than by calling 911. Software would aggregate those reports into a constantly updated map of the situation, which citizens and emergency responders could check without encountering clogged phone lines. People could also register to receive block-by-block information — again via e-mail or text message — about whether to stay in place, to evacuate or to respond in some other way. And people would be able to coordinate with their neighbours before, during and after emergencies on community message boards. So, for example, a family could agree to take responsibility for a wheelchair-bound neighbour during an evacuation.

S. B. LIU

Researchers tracked social norms in the communications following the Virginia Tech shootings in 2007.

F. J. BROWN/AFP/GETTY IMAGES

Mobile-phone photos of the Sichuan earthquake in China aided the rescue efforts.

Shneiderman is collaborating with Wu, Qu and others to explore how a small-scale version of the system would work on the University of Maryland campus. Wu says that the official emergency responders in the university Department of Public Safety are very interested in the idea — but they are also wary. They are concerned that they wouldn't have the staff to run the system, he says, and that it might confuse people who are already well-trained to call 911 in an emergency. Then there's the possibility that people could use the system to spread bogus information and rumours. "If you have everybody able to do peer-to-peer, then there is widespread panic," one safety officer told the researchers.

That reaction — interest tinged with scepticism — mirrors the reaction of many professional emergency responders.

Federal and local disaster-response agencies have long operated under the Incident Command System, a standardized protocol developed in the 1970s by firefighters battling California wildfires, and later adopted by federal agencies including the Federal Emergency Management Agency (FEMA).

The system, with a clear, top-down chain of command, views communication with the public as a one-way street: information is supposed to flow from officials to the public via warnings sent out over TV, radio and other media. That view fits well with older academic research in disaster sociology.

"Thirty years ago no one talked about crisis communication," says Dennis Wenger, a sociologist at the National Science Foundation who has studied disasters for more than three decades. "Back then the term was warnings research."

The questions that preoccupied researchers then were how to write an effective warning — instructing people to evacuate before a hurricane, for example — and how best to make sure everyone heard it.

"But then all of a sudden came the Internet revolution, and it blew apart this notion of a linear chain," says sociologist Kathleen Tierney, director of the Natural Hazards Center at the University of Colorado. Social-networking and photo-sharing sites, mobile phones and text messages have turned the chain into a web.

In hindsight, that was only to be expected. The chain was never as linear as the models made it out to be. One of the first to realize that was Thomas Drabek, a disaster researcher at the University of Denver, Colorado, who surveyed evacuees after the South Platte River flooded near Denver in 1965. More than 60% of the people he spoke to told him that even after they received a warning telling them to evacuate, they tried to confirm it — checking with family and friends, talking things over, watching to see what their neighbours were doing — before taking action.

Reaching an understanding

Researchers have observed the same information-sharing inclination in many disasters since. After the bombing of the World Trade Center in New York in 1993, Benigno Aguirre, a sociologist at the Disaster Research Center of the University of Delaware in Newark, found that people who were in large offices took longer to evacuate than people who were in smaller groups, because it took them longer to come to a common understanding of what was happening.

> **"How do you convince people that they are at risk? Only through other people."**
> — Dennis Mileti

"It is actually very difficult to get human beings to perceive that they are at risk," says Dennis Mileti, a disaster-management researcher at the University of Colorado. "How do you convince people that they are at risk? Only through other people." And there, as long as the process isn't too time-consuming, is one of the advantages of the Internet. "Long before technology, you could check in with neighbours next door," says Sutton. "But now you can check in with peers around the country."

Garry Briese — a regional administrator for FEMA in Denver and, for 22 years before that, the executive director of the International Association of Fire Chiefs — is excited about the opportunity that public feedback offers. He recounts the example of a wildfire in Deckers, Colorado, last summer, when he was able to find photos of the fire posted online by local residents within 30 minutes of hearing about the fire himself.

"It expands our ability to have situational awareness," he says. "Someday, in a fast-moving wildfire, if we had the right system, we could ask people by a text alert to take pictures and send them to an emergency centre. And our ability to understand what the scene looks like would be enhanced."

That scenario is still a long way from reality, says Briese, who, in July 2008, talked to an official from a large city's emergency operations centre about the information already available on the web. "A few days later, the guy called me and said 'you know, my IT department has me blocked from those sites'," he says. "So here we have an emergency manager who wants to learn, and the first thing he has to do is go to the IT department and convince them."

Sutton encountered something similar when she spoke about her research at a conference of state-level emergency managers in September 2008. "I had a couple of people say to me, 'This is so new to us, we don't even know what to ask'," she says.

In any case, answers might not have been easy to supply. "This information is out there, and you can find it online," Sutton says. "But how do you aggregate it, and where would it get plugged into? Would it go back to an emergency operations centre, a joint information centre, a public information office?"

When Shneiderman proposed community response grids in 2007, he thought that it might become a reality within 3–5 years. Now, with so much research still to do, he thinks 5–10 years might be more accurate. But he is used to waiting. He helped to develop the concept of hypertext links in the 1980s and saw them become integral to the web more than a decade later.

Not everyone has Internet access or the technical knowhow to take advantage of it. But, during a disaster, a community-response grid could benefit almost everyone, as family, friends, neighbours and authorities share what they know. "I think this is inevitable," says Shneiderman. ■

Lea Winerman is science editor for *The Online NewsHour with Jim Lehrer*.

CORRESPONDENCE

Scientists stand by decision to join Mbeki's AIDS panel

SIR — Your Editorial 'The cost of silence?' (*Nature* **456,** 545; 2008) questions our decision — as scientists who opposed dissident theories — to participate in the now-discredited AIDS advisory panel set up by former South African president Thabo Mbeki.

We had no role in, and did not see or approve, the panel's report (*Nature* **410,** 730; 2001). Each of us had agreed independently to join the panel, guided by our consciences as scientists in a young democracy and without prior knowledge of who else had been invited to participate. To us, it provided an opportunity to present Mbeki with the alternative viewpoint and the compelling scientific evidence that HIV causes AIDS, in the hope that rationality would prevail.

However, Mbeki's antipathy to antiretroviral drugs was influenced by documents from and interactions with AIDS dissidents that predated the setting up of the panel. We underestimated the strength of his dissident views on AIDS and how little impact sound science would eventually make on them.

Sadly, our advice to Mbeki on AIDS causation and antiretroviral treatment was rejected. We cannot, therefore, be numbered among those held accountable for Mbeki's decisions, which led to the loss of many thousands of lives in South Africa through lack of access to antiretroviral therapy.

That we failed to change Mbeki's opinions on AIDS is a matter of record. But your Editorial is unreasonable in implying — with the benefit of hindsight — that scientists could have foreseen this failure and therefore should not have signed up to an opportunity to give the president critically important information that might have saved the lives of their fellow-countrymen.

We stand by our decision to participate in the Mbeki panel. We have an obligation to our country, which is suffering the worst AIDS epidemic in the world, to do everything in our power to provide our political decision-makers with the best scientific advice, whether or not they are a priori opposed to or supportive of our views.

Salim S. Abdool Karim, Hoosen M. Coovadia, Malegapuru W. Makgoba
CAPRISA, Doris Duke Medical Research Institute, Nelson R. Mandela School of Medicine, University of KwaZulu-Natal, Private Bag X7, Congella 4013, South Africa
e-mail: karims1@ukzn.ac.za

When winning a Nobel prize seems to run in the family

SIR — In his Correspondence 'You're the best man for this job, son. What a coincidence!', Albert Ruggi's suspicions about the process by which the offspring of professors are deemed to be the best candidates for new positions may well be justified (*Nature* **456,** 870; 2008). On the other hand, a few rare families just do produce generations of eminent scientists. For example, there are at least seven parent–child pairs of Nobel laureates.

Four of these were in physics: the Thomsons (J. J. in 1906 and George in 1937), Braggs (William and Lawrence together in 1915), Bohrs (Niels in 1922 and his son Aage in 1975) and Siegbahns (Manne in 1924 and his son Kai in 1981). Marie Curie and her daughter Irène Joliot-Curie both won the Nobel Prize in Chemistry (1911 and 1935), after Marie and her husband, Pierre, had won the physics Nobel in 1903.

The Kornbergs branched out more (Arthur, physiology or medicine, 1959; Roger, chemistry, 2006), as did Hans von Euler-Chelpin (chemistry, 1929) and his son Ulf von Euler (physiology or medicine, 1970).

Jay M. Pasachoff California Institute of Technology 150-21, Pasadena, California 91125, USA
e-mail: jmp@caltech.edu

Ecologists should join astronomers to oppose light pollution

SIR — In his Commentary 'Time to turn off the lights' (*Nature* **457,** 27; 2009), astronomer Malcolm Smith argues for darker skies. Ecologists would also do well to support the International Year of Astronomy, considering the potentially severe impact of light pollution on some biological systems. Moths, for example, as well as the bird migrations Smith mentions, are adversely affected by light pollution.

In most instances, the origin of atmospheric and ecological light pollution is identical. So ecologists should back initiatives stimulated this year by physicists, to help raise awareness of the undesirability of light pollution across different disciplines.

Josef Settele UFZ, Helmholtz Centre for Environmental Research, Theodor-Lieser-Strasse 4, 06120 Halle, Germany
e-mail: Josef.Settele@ufz.de

Lindauer's genius showed evolution in a simple experiment

SIR — I was saddened to learn of the death of Martin Lindauer, an under-appreciated hero of science. Thomas D. Seeley, in his Obituary (*Nature* **456,** 718; 2008), describes experiments from Lindauer's *Communication Among Social Bees* (Harvard Univ. Press, 1961) that demonstrate his talent. The book also includes experiments that possess what physicist I. I. Rabi used to call 'witz': an unexpected twist that elevates an experiment to a higher level.

Living in an enclosed, sheltered space, the honeybee *Apis mellifera* performs its communicative dance on a vertical surface in the dark, using gravity as a substitute for the direction of the Sun. By depriving them of a vertical surface and giving them a direct view of the Sun, Lindauer forced them to revert to the more primitive, Sun-directed dance of their dwarf Indian relative, *Apis florea*.

In closing the gap between a primitive and an advanced condition, Lindauer possibly produced the best-ever experimental evidence for evolution. Scientists concerned with evolution of human language and mind might ponder his success.

William L. Abler Department of Geology, The Field Museum, 1400 South Lake Shore Drive, Chicago, Illinois 60605, USA
e-mail: wabler@fieldmuseum.org

Culture clash in Chinese university: a response

SIR — In your Editoral 'Culture clash in China' (*Nature* **456,** 545–546; 2008), you incorrectly say that I am professor emeritus, having retired from the College of Life Sciences, Peking University, four years ago. In fact, I retired in February 2006 and do not have emeritus status. Neither did I retain my laboratory there in order for my associate professor to take it over formally as a way of maintaining my influence.

I have kept my laboratory running with the help of a grant from the National Natural Science Foundation of China (NSFC). When I applied to the NSFC, Peking University guaranteed my lab and equipment until I had completed the work. The associate professor you mention was a co-author on this grant application.

Although I did submit an online posting accusing Yi Rao, the dean of life sciences, of withdrawing the laboratory for use in other applications (http://tinyurl.com/8l4u9x), I have never proposed that the associate professor should take it over. My aim is that he should be able to use it to continue his research.

Keming Cui College of Life Sciences, Peking University, Beijing 100871, People's Republic of China
e-mail: ckm@pku.edu.cn

ESSAY

Kinship: Race relations

Our notions of family, population and race may need revising in the age of personal genomics, argues **Aravinda Chakravarti**.

Genealogical records are currently the system of choice for people tracing their family history. But in the next decade, we will be able to identify many of our relatives by searching a DNA database of personal genome sequences. There are good reasons for switching to DNA: in general, historical records cover at most the past 500 years; our genomes, in contrast, bear the stamp of tens, if not hundreds of thousands of years of history. Even individuals without genealogical records will be able to correctly create a family tree with connections to known relatives, to those they were unaware of, and to relatives so distant that they stretch the meaning of the word 'family'.

Two developments are making it possible for geneticists to begin homing in on the patterns of relatedness between the world's seemingly diverse people. The first is the discovery that we share very recent common ancestors. Until the late 1980s, our early hominin ancestors, from a few million years ago, had been found in many locations, and modern humans were thought to have arisen from the local evolution of these species in different parts of the world.

In 1987, Allan Wilson and his students presented a new and surprising human history: their comparison of only a fraction of our genome — that found in the mitochondria — in a handful of Africans, Asians and Europeans, showed that all living humans are related via a set of common ancestors who lived in Africa about 200,000 years ago. Other studies have since shown that the world beyond Africa was settled even more recently. From 100,000 years ago, descendents of our African forebears spread out to populate other continents (the New World, perhaps as recently as 25,000 years ago), with the lineages from different settler groups eventually mixing through further migration.

The striking implication of this is that all living humans are mosaics with ancestry from the many parts of the globe through which our ancestors trekked. In other words, each of us has around 6.7 billion relatives.

The second change that is allowing geneticists to piece together human ancestry is remarkable progress in our ability to study DNA. Driven mainly by the desire to find genetic traits that underpin common diseases, extraordinarily rapid technological advances in DNA and computing analysis are allowing geneticists to compare more than a million markers, or variable DNA sites, in people's genomes, and to make such comparisons between hundreds of thousands of people.

> "The global picture of relatedness that is emerging from DNA studies stands to shatter many of our beliefs about ourselves."

The global picture of relatedness that is emerging from DNA studies, in the context of established facts about our recent common ancestry, stands to shatter many of our beliefs about ourselves. In particular, it calls into question existing ideas about populations and race.

Extended families

Anthropologists and sociologists have conventionally assessed kinship by asking people about their social relationships with others. This provides a notoriously incomplete, perhaps even erroneous, picture of biological relatedness for living humans, and little more than myth or speculation when used to assess the connections between assumed forebears. As the saying goes, "Maternity is a matter of fact, paternity a matter of opinion." In the United States, findings from organizations that assess paternity, for example to persuade fathers to support their disputed yet biological children, suggest that at least 1 in 20 people don't know the identity of their genetic father. Also, in the case of isolated people such as the Old Order Amish, all individuals share a small group of common founder ancestors. So a social relationship, such as first cousin or parent, is an imperfect guide to their genetic relatedness.

The only way to assess biological kinship with any certainty is to look for the stories of ancestry marked indelibly in a person's DNA.

In the past decade, technological advances have reduced the costs of examining entire human genomes 1,000-fold or more. These have largely been driven by a desire to identify the genes underlying common chronic diseases or adverse drug reactions. Already, more than a million marker sites in the human genome, have been examined in some 100,000 people to identify more than 300 novel common disease factors. And, increasingly, researchers are sequencing all 3 billion DNA base pairs of the human genome; one international study, the 1,000 Genomes Project, aims to sequence the genomes of 1,000 people over the next two years.

However, the interactions between myriad genetic and environmental factors seeming to underpin most common diseases are proving to be highly complex. As a result, large-scale comparisons of people's DNA may be bringing geneticists closer to understanding how the world's people are related to one another, than to establishing the causes of common diseases.

DNA studies scour the genomes of 'unrelated' patients for common genetic patterns that are absent in other 'unrelated' people without the disease. For example, my

G. BECKER

colleagues and I have examined the genomes of 16,000 people to identify a region on chromosome 1 that harbours a gene affecting the risk of sudden cardiac death (in combination with many other genetic and environmental factors). But our finding that markers in this region are over-represented among patients compared with those who are disease-free indicates that the disease arises, in part, from shared ancestry. In other words, the disease runs in families, even though the family links may be thousands of years old.

Even putting aside the disease factor, we can uncover both the proximal and remote ancestral relationships of any two of these 16,000 people by comparing the degree and pattern of similarity across their genomes. Indeed, comparisons of millions of markers, and certainly of entire genomes, will identify far more specific relationships between strangers than has been uncovered by the ancestry tests now in vogue.

Increasingly, customers pay companies to convert their DNA into ancestry information. But most if not all such pictures of relatedness are based on markers on the mitochondrial genome (inherited only from mothers) or on the Y chromosome (inherited only from fathers). These represent a minute fraction of our genetic inheritance (less than 1%), so give a highly incomplete picture of relatedness. Companies also tend to compare people's DNA to sequences held in their own private databases, which are currently too small to uncover more than the continental origin of a person's ancestors.

It is not inconceivable that studies investigating the genetic basis of diseases will reveal people's previously unknown cousins, siblings, half-siblings or even parents. Human geneticists are bound by consent forms not to reveal cases of mistaken paternity if they discover them. But if databases of DNA sequence information become publicly available, just as genealogical records are now, people will be able to compare their own genome sequence with those of millions of others. Children born to mothers artificially inseminated by an anonymous donor could potentially discover their numerous half-siblings. Even in the case of remote relationships, people may interact, perhaps through online social networking, with newly found, distant relatives regardless of their culture, politics and race. Such a scenario is increasingly plausible given people's willingness to share personal information online, for example in social-networking websites.

Only skin deep

Perhaps the most striking consequence of more and more people having their entire genome examined for genetic variation is the blurring of our concept of discrete human populations. Current thinking, championed by anthropologists and buttressed by old genetic data, is that human populations are intact groups that have had their own language and culture for eons. In fact, the population is thought to define an individual's genetic identity, and kinship between individuals is considered only within the context of this or that group. It's within this context that people trace their ancestry using genealogies or ancestry tests, and that the discovery that President Barack Obama is related to Dick Cheney makes news.

> **"We are all multiracial, related to each other only to a greater or lesser extent."**

Currently, the population view dominates in genetics because researchers sample clusters of individuals from distantly related groups. The clearly observable, or measurable, physical and genetic differences between people are especially marked when people from the peripheries of the spectrum of human variation are compared — so, for example, when Africans are compared with Europeans or Asians.

Race has long been a socio-political construct. But by focusing on the effects of natural selection on genes whose effects are visible, and sampling people from the extremes of human diversity, geneticists have unwittingly (and sometimes wittingly) added credence to society's views on separateness by genetically characterizing racial categories.

However, the current picture emerging from genetic studies is that we are all multiracial, related to each other only to a greater or lesser extent. More detailed data on genetic variation, along with an improved sampling of humanity, are showing continuity in variation across the globe, not abrupt transitions between population-specific sequence patterns. Differential population growth, about 10,000 years ago, based on the evolution of agriculture, technology and politics seems to have made sparse isolates of our species into the 'major' groups of today. In other words, except for immigrants, kinship between two humans seems to be directly related to the geographical distance between their birthplaces.

An even clearer, and unbiased, picture of humanity's genetic diversity and relationships would emerge if geneticists focused on individuals instead of populations. This may involve sampling humans randomly across a grid, and then assessing their individual and group features (such as birth place, parental birth places, language and group affiliations). Genome-wide studies carried out in this way could result in individual identity and kinship coming to define populations rather than the other way around. We could test once and for all whether genetic race is a credible concept.

This would be tremendously exciting. It is bound to stir up our deeply held notions of who we are, where we came from, our history and thus our politics. More often than not, the views of society have shaped science rather than the other way around. In this instance, it may be time for science to reshape the views of society. By dismantling our notions of race and population, we may better appreciate our common, shared and recent history, and perhaps more importantly, our shared future. Overhauling such concepts in the light of genetic research is particularly important if we are to accommodate the changing face of human groups around the world thanks to increased immigration to distant lands. Such migration has happened before, as our genomes show, only slowly, over 150,000 years. ■

Aravinda Chakravarti is at the McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Broadway Research Building, Room 579, 733 North Broadway, Baltimore, Maryland 21205, USA.
e-mail: aravinda@jhmi.edu

See http://tinyurl.com/8ub4f4 for further reading. For more on Being Human, see www.nature.com/nature/focus/beinghuman.

BOOKS & ARTS

Evolution's challenge to genetics

Do conjoined twins and two-legged goats suggest a minor role for genetics in evolution? The evidence is not strong enough to upset the orthodox view, argues **Jerry A. Coyne**.

Freaks of Nature: What Anomalies Tell Us About Development and Evolution
by Mark S. Blumberg
Oxford University Press: 2009. 344 pp.
£12.99, $22.95


Darwin200

In 1980, an evolution meeting at the University of Sussex, UK, featured several speakers who questioned the importance of genetics in understanding evolution and development. The 'structuralists' saw the adaptations of many organisms as products of self-organizing molecules rather than natural selection. Others touted the epigenetic view, claiming that important evolutionary change involved heritable features not coded in the DNA. Perturbed, the distinguished embryologist Lewis Wolpert stood up and proclaimed that he too had a radical and heretical view: "Genes control development." Wolpert was puckishly defending what he saw as a perfectly adequate paradigm against those who minimized the importance of genes. To emphasize his point, he switched the lights on and off during the coffee break — but the structuralists refused to admit that the switch controlled the lights.

The past three decades have vindicated Wolpert. Virtually all of the major advances in evolutionary developmental biology, or 'evo-devo', have been firmly grounded in genetics. These include studies — two of them awarded a Nobel prize — on how genes organize body plans, how genes are regulated and how the same genes, such as *PAX6*, are recruited in the independent evolution of similar structures in different species.

But as evo-devo blossoms, the anti-genetics strain persists. Evo-devo is undergoing the kind of spasm experienced by palaeontology when Stephen Jay Gould and others decided that their field had been overlooked in the modern evolutionary synthesis, or worse, forced to conform to the theories of population geneticists and others at the heart of the synthesis. Gould suggested that palaeontology, when uncoupled from its overbearing cousins, would provoke a major revision of our understanding of the evolutionary process. Gould was wrong: the neo-Darwinian perspective emerged unscathed from its encounter with punctuated equilibrium. Now it is the turn of evo-devo to challenge neo-Darwinian orthodoxy. As Mark Blumberg declares in *Freaks of Nature*, evolutionary biology should not just describe developmental phenomena, but should also incorporate new evolutionary processes that de-emphasize the role of genetics.


UPPA/PHOTOSHOT

It is difficult to see how developmental quirks are relevant to evolution — even if you have two heads.

By presenting a parade of animal 'freaks' — mutants, developmental anomalies and weird species — Blumberg imparts lessons that, although familiar to biologists, will be valuable to non-specialists. He emphasizes that the complex process of development can be unravelled by understanding how such anomalies are produced. Conjoined twins, for example, tell us something about how identical twins form. He highlights, correctly, that there is no direct correspondence between genes and traits: there is no such thing, for example, as the gene for the thumb, although some mutations can create new thumbs. He also stresses that evolution can work only by modifying development, and that natural selection can create individuals whose development responds adaptively to the environment — as in the case of rotifers, some species of which develop spines in response to predators.

Blumberg illustrates his points with clear and intriguing examples. We learn that female hyenas lack vaginas but have huge, penis-sized clitorises through which they copulate and give birth, often resulting in high infant mortality. We meet Abigail and Brittany Hensel, 18-year-old conjoined twins from Minnesota — two heads on a single body — whose touching story can be seen on YouTube (see http://tinyurl.com/26bpm9). And we discover that the nervous system of rabbits can be entrained to make them walk rather than hop, implying that hopping is not genetically encoded but a by-product of the rabbit's leg structure.

Blumberg's ambitions transcend storytelling: he aims to show that developmental biology has made real contributions to evolutionary theory. The theory's problem, as Blumberg maintains, is its "gene-centered, population-level thinking", also described as "simplistic single-cause, gene-centered thinking". What paradigms, then, should supplant our misguided embrace of Gregor Mendel?

The first is epigenetics. Blumberg notes that larger male dung beetles roll larger balls of dung, which in turn nurture larger sons. He argues that "Such examples of nongenetic transmission of characters are now becoming commonplace and are helping solidify the notion that the heredity upon which evolution depends is more than just about genes." But we must be careful. Some adaptive 'epigenetic' phenomena, such as parental imprinting of chromosomes, which influences gene expression depending on which parent passed on the gene, are based on instructions in DNA. Other

cases of epigenesis, such as the conformational changes of prion proteins, are of minor evolutionary significance. Still others, such as an ancestral cell's ingestion of the bacteria that evolved into mitochondria, were of immense importance in evolution but are infinitely rarer than adaptive changes based on genes. And in nearly all cases, epigenetic effects peter out after a few generations, unable to promote major evolutionary change. Perhaps the most serious criticism of epigenetics is that of the thousands of inherited mutations found in model organisms such as mice and fruit flies, virtually all reside in DNA.

The second paradigm involves "the self-righting properties of developmental systems", also known as phenotypic accommodation, which, says Blumberg, lead to evolutionary innovation. To support this, he trots out the two-legged goat described in 1942 by Dutch veterinarian E. J. Slijper. The goat, a developmental anomaly born without forelegs, learned to hop on its hindlimbs like a kangaroo. When the goat died in an accident — some say in an ill-advised experiment to see if it could walk downstairs — Slijper's autopsy showed that it had undergone modifications of its spine, hindlimbs, muscles and neck that facilitated its bipedal hopping. It has been suggested that this hobbled beast is a model for the origination of bipedality in some lineages, even humans. Although Blumberg admits that this is unlikely — after all, there are perfectly good, and more parsimonious, Darwinian explanations for bipedality — he approvingly quotes anatomist Pere Alberch: "The regulatory capacities of an epigenetic system imply that any intrinsic change will trigger a sequence of regulatory changes to automatically generate an integrated phenotype." But as this integrated phenotype was not based on genetic differences from any other goat, it could not be transmitted to offspring, and its relevance to evolution is unclear. The phenotypic changes in Slijper's goat did not result from some inherent self-regulating property of development. Rather, they reflect an evolved phenomenon: natural selection has given bones and muscles the adaptive property of developing in response to the stresses they experience.

Blumberg's final alternative to conventional evolution is genetic assimilation. As with phenotypic accommodation, here the phenotype changes before the genes. During assimilation, an initial environmental change alters the phenotype of many individuals, exposing previously hidden genetic variation that can then be selected. Eventually, what was an environmental change becomes genetic, mimicking the inheritance of acquired traits.

Social learning is one way to start this process. For example, in the 1920s, two species of British tits learned by mass imitation to pry up the foil on milk bottles and drink the cream on top. Were home milk delivery still common, one can imagine that this propensity might have become genetically assimilated. Individuals with greater abilities to learn the behaviour, or perform the actions needed, would be favoured by selection. Eventually, cream pilfering would become innate — coded in the genes. Many adaptations might have started in this way; fish, for example, may have evolved adaptations for living on land after some individuals discovered terrestrial insects to be a rich food source. But we can also explain such cases by invoking simple selection on pre-existing genetic variation. In the absence of a single credible example of genetic assimilation in nature, it remains an appealing but untested speculation.

In the end, the problem with these explanations is not so much that they are wrong, or of no potential importance in evolution. Rather, it is that Blumberg gives the impression that they are established truths rather than hypotheses that have remained unconfirmed for three decades. In his anxiety to boost the status of evo-devo in the pantheon of evolutionary subdisciplines, Blumberg has short-changed orthodoxy. Not only does the traditional view of evolution explain far more than he allows, but Blumberg shapes his own vision of development to inflate its challenge to neo-Darwinism. I, for one, am with Wolpert. ■

Jerry A. Coyne is a professor in the Department of Ecology and Evolution at the University of Chicago, Illinois 60637, USA.
e-mail: j-coyne@uchicago.edu

The future is now

"With great power comes great responsibility," uttered Stan Lee's comic-book superhero Spider-Man in his first published appearance in *Amazing Fantasy* in 1962. Since then, science has advanced to such a point that the human body can be enhanced in ways that mimic fiction. Cloning, face transplants, prosthetic limbs, brain–machine interfaces and cognition-enhancing drugs promise utopian or Orwellian visions of the future, depending on your outlook. It is the scientific community, not the superhero, that holds the great responsibilities of our age and the next.

Two books, *Human Futures* and *The Science of Heroes*, use a crystal ball to imagine how science will determine the future of human existence and society.

Human Futures is a varied collection of meditations on the notion of humanity from researchers, artists, philosophers and even a Blood Elf priest from the online role-playing game *World of Warcraft*. The book is born of the Human Futures programme of the Foundation for Art and Creative Technology, based in Liverpool, UK. The programme comprised a series of exhibitions and lectures for public debate, in which creative assemblages of artists, philosophers and scientists explored questions of what we are now, and what we will become.

The first of the book's four sections, entitled 'Visions', perhaps best succeeds at describing the problems of the future using intelligent insights from the present. Many of the essays focus on society's willingness, or not, to embrace new technologies. Philosopher Russell Blackford argues convincingly that a fear of new technologies is groundless and erodes liberalism. He notes that a governing state can allow new advances, such as choosing the sex of one's child, without endorsing the underlying technology or morality of an individual's choice.

The concept of human enhancement is investigated by ethicist Ruud Ter Meulen. Certain drugs such as modafinil, used to treat narcolepsy, have been shown to improve cognition and are becoming increasingly popular with students revising for exams. If everyone takes the drug in future, then what is the new norm? Will it make us better, or just different?

Physicist Richard Jones asks how humans have been enhanced by technology. He observes that the public accepts the benefits of gadgets while increasingly rejecting the scientific world view. As a result, he explains, per capita energy use in the United Kingdom has risen from 20 gigajoules per person in 1800 to nine times that figure today. Life expectancy, he notes, is strongly correlated with energy use. Such a rise in energy use looks unsustainable at present, but Jones asserts that technology is a product of society, not a runaway automaton, and solutions will come as long as the energy flows.

Oddly, the section concludes with a delightful

Human Futures: Art in an Age of Uncertainty
Edited by Andy Miah
Liverpool Univ. Press: 2008. 368 pp. £35

The Science of Heroes: The Real-Life Possibilities Behind the Hit TV Show
by Yvonne Carts-Powell
Berkley Boulevard Books: 2008. 288 pp. $15

NBC UNIVERSAL PHOTO

Many of the 'superpowers' portrayed in the television series *Heroes* are no longer science fiction.

but misplaced essay on 'evidence dolls', the creation of designers Anthony Dunne and Fiona Raby. The plastic miniatures are a hypothetical future product in which to store sperm and hair samples of prospective reproductive partners. The dolls are personified by four women, who reveal how knowledge of their partner's genes might influence their sexual and reproductive lifestyles.

Two other essays that centre on human enhancements — such as the lower-leg prostheses that allow South African runner Oscar Pistorius to compete with able-bodied athletes — are left to later sections. This misfiling of the book's content makes its arguments and themes hard to follow. Sociologist Steve Fuller's history of humans playing God is a thoughtful reminder of how badly we have handled aspects of divinity in the past, but it belongs nearer to Pramod Nayar's narrative on post-human rights. Both essays grapple with the prospect of endowing legal and moral status to our cybernetic or genetically enhanced descendants.

The book's over-designed layout — a hybrid of uber-stylish photography mixed with elements of the record sleeve from Radiohead's seminal album *OK Computer* — does not aid the reader experience. Indeed, the voice of musicians is absent, which weakens the book's proclamations of diversity.

The Science of Heroes explores similar themes, but in a very different style. It uses the vehicle of *Heroes*, the popular sci-fi television series that follows the moral chaos inflicted on a clutch of people who have incredible powers proportional to their youth and good looks. In the book, Yvonne Carts-Powell propels the reader on an enthusiastic and entertaining journey through the realms of biology and physics that might one day produce the first genuine super people.

And that day might come sooner than we think. The regenerative powers of *Heroes* character Claire are ones we already possess — we are just slower to heal. Hiro's teleportation has been achieved, at least in the subatomic world, and metamaterials with negative refractive indexes promise invisibility for all. Other powers, such as the ability to steal memory, are already with us in the form of drugs that help victims of trauma to forget.

Historically, superheroes are a snapshot of the relationship between science and society at any one time, based on their powers and how they obtained them. From the perfect Superman born of a 1930s United States adopting eugenic practices, to the nuclear-powered Spider-Man of the cold-war era and the psychologically disturbed Batman incarnations of the therapy-obsessed 1980s and 1990s, every generation has its super people. Those portrayed in *Heroes* are the genetically enhanced Generation X of the superhero world — too busy to save the world because their own lives are in perpetual turmoil.

Societies get the superheroes they need most. *Human Futures* and *The Science of Heroes* give a tantalizing glimpse of how science might make this a reality in a future human existence. And given Carts-Powell's talent for explaining the science, perhaps an army of her clones will take over high-school education. ■

Arran Frood is a science writer based in the United Kingdom.
e-mail: arranfrood@gmail.com

Is there life on Europa?

Unmasking Europa: The Search for Life on Jupiter's Ocean Moon
by Richard Greenberg
Praxis/Springer: 2008. 278 pp.
£17.50/$27.50

In the field of astrobiology, the discovery of life beyond Earth sits like a gem inside the nested Russian dolls of physics, geology, chemistry and, ultimately, biology. Efforts to understand the habitability of worlds within our Solar System began with physical and astronomical surveys, and have now moved on to the challenge of cracking open the geological secrets of key destinations such as Mars and the large, icy moons of Jupiter and Saturn.

Understanding the geological context for life is critical. Rock cycles, whether they are of silicates or ices, enable chemical cycles that can then be exploited by biological systems. Such cycles are central to life on Earth. On Mars, the demise of mantle convection may have led to the planet becoming cold and dry. Near the giant planets of the outer Solar System, and perhaps around massive extrasolar planets, rock cycles may be driven by the gravitational squeezing of icy moons due to tidal interactions. On icy moons such as Jupiter's Europa, the mixing of irradiated, oxidant-rich surface ice with a water ocean could maintain a chemically rich environment capable of sustaining life.

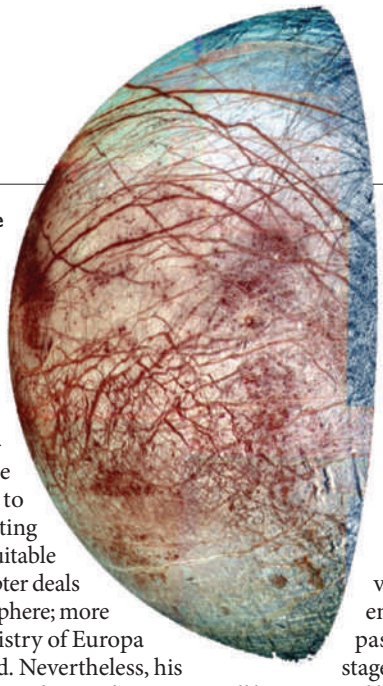
In *Unmasking Europa*, planetary scientist Richard Greenberg details in depth our geological understanding of the tidally tormented icy surface of Europa. Without pulling any punches, he also describes the equally tormented scientific debate that has led to the current canon. More than a decade after the Galileo spacecraft returned magnetic-field and gravity data that strengthened the case for a subsurface, liquid-water ocean on Europa, we still do not know whether that ocean lies beneath an ice shell just a few kilometres thick or a shell with a thickness of more than ten kilometres.

From an astrobiology perspective, a thin shell could permit direct cycling of oxidant-rich ices with the ocean. A thick ice shell, however, would impede the cycling of surface material, possibly limiting the chemical energy available to any life below the surface. On this contentious debate over the ice thickness, Greenberg notes, "by itself, modelling of heat transport on Europa is too uncertain to definitively discriminate between thin conductive or thick convective ice". However, on the basis of a host of geological features observed in images from the Voyager and Galileo missions, many of which are reproduced in the book, Greenberg argues compellingly that only a thin shell is consistent with the observed ridges, cycloidal features and chaotic terrain of Europa, all of which can be explained through tidal dynamics.

Although Greenberg occasionally strikes an acerbic tone when describing scientific differences with those on what he calls the thick-ice bandwagon, his motivation seems noble. He fears that "the most brilliant young minds

may leave science if they perceive it to reward something other than good research". He feels that the data point towards a thin ice shell but that political powers have marginalized this interpretation and those scientists who advocate it. When discussing his own work, Greenberg generously bestows much credit on his former students, postdocs and colleagues.

In *Unmasking Europa*, Greenberg succeeds in conveying a story, not of heroes and villains, but about the rise and fall of ideas and how some become accepted for reasons that perhaps go beyond empirical support. In Greenberg's earlier work, *Europa the Ocean Moon* (Springer, 2005), which is of similar scope but targeted to a research audience, the political storyline is not particularly appropriate. In his latest work, he delivers an accessible and well-laid-out popular-science treatment in which the political narrative is more pertinent, although obviously biased towards his own perspective. Greenberg uses humour to balance out the tone, as in his suggestion that the reader should buy a second copy of the book just to cut out the images and do the geological reconstructions while reading the first copy.



NASA/JPL-CALTECH

Europa's fractured icy surface could conceal life beneath.

Tides are the recurring theme of Greenberg's treatment — they "connect the orbits of Jupiter's moons to the geology of Europa, creating environments potentially suitable for life". Only one short chapter deals with the possibility of a biosphere; more detail on the known chemistry of Europa would have been welcomed. Nevertheless, his treatment of tidal dynamics is thorough.

Europa has not yet revealed a smoking gun, as have the icy plumes of Enceladus, to indicate that it is geologically active today. This has left the planetary geology community staring at the limited imagery of Europa, wondering what its surface features reveal about the interior. Centuries ago, geologists began adopting the uniformitarian mantra of 'the present being the key to the past'. In the ebb and flow of planetary science, with data streams punctuated by missions that are all too rare, we often find ourselves struggling to decipher the geological present, much less the past.

Unmasking Europa provides a comprehensive and engaging account of Europa's past and present, and sets the stage for the many questions that will be answered by future missions as we continue our search for life beyond Earth. ■

Kevin P. Hand is a scientist at NASA's Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, California 91109, USA.
e-mail: khand@jpl.nasa.gov

Scripting scientists' lives

Leave a Light On
Ensemble Studio Theatre, New York City
22 January 2009. Part of the First Light Festival, which runs until March 2009.

Last year, at a first reading of her play about the life of biologist Robert Trivers, Ann Marie Healy noticed a stranger in the back of the theatre, laughing. Afterwards, the man strode over to the actor who had played the young biologist as a foul-mouthed and promiscuous genius working out the evolutionary logic of human kindness and conflict, and said: "You got it exactly right." That stranger was Trivers.

Healy's play features in New York's First Light Festival, a collaboration between the Ensemble Studio Theatre and the Alfred P. Sloan Foundation that incubates science-based theatre. The festival, which has run annually for more than a decade, includes nine full-length plays this year and continues until the end of March.

In *Leave a Light On*, Trivers is portrayed as an ambitious, untenured professor who ruffles feathers at Harvard University's department of zoology as he attempts to take a Darwinian approach to human nature. Dissatisfied with an academic culture that is hostile to his ideas, Trivers retreats to Jamaica to study lizards and then moves to teach at the University of California, Santa Cruz, where he meets Huey Newton, former leader of the Black Panther party. Widely believed to have dropped 'off the grid', Trivers returns to academia more than a decade later to study the adaptive value of self-deception.

Healy weaves in the science with a light touch. In the play, with the help of a female colleague who is also a love interest, Trivers works out his theory of reciprocal altruism using a series of imaginary birds with distinct approaches to selfless behaviour: Suckers, who always groom their peers; Cheaters, who never do; and Grudgers, who only groom tit-for-tat. As in Tom Stoppard's play *Arcadia*, the script darts between centuries and characters, punctuating Trivers's sobering career with farcical episodes from the courtship of Charles and Emma Darwin that are meant to explain the logic of gene competition.

R. RESSMEYER/CORBIS

Robert Trivers's ideas on behaviour caused conflict.

The play hardly needs such asides: Trivers's own ideas are enough to drive the plot.

The 2009 First Light Festival began with an uneven selection of one-act plays collectively called $E = mc^{brunch}$, portraying a chemist discovering her brother's meth lab, an Olympic gymnast trying to prove her rivals are underage, and a mathematician confronting risk in an airport restaurant. The full-length plays take on an equally wide range of topics. Anna Ziegler's *Photograph 51* portrays the familiar story of biophysicist Rosalind Franklin, whose X-ray diffraction images led the way to the discovery of DNA structure in 1953. Tommy Smith's *Beautiful Night* will show Soviet inventor and electronic-music pioneer Léon Theremin falling in love with a black ballerina in New York City in the 1930s — with live accompaniment from the eerie-sounding theremin instrument. And in the improbable monologue *Five Easy Steps to Metaphysical Fitness: They Actually Work*, comedian Emily Levine will impart wisdom gained by staging a one-woman show about physics while struggling with her pituitary-gland disorder.

"The goal is not just to demystify science but to show its intrinsic appeal, both emotional and intellectual," says Darcy Kelley, a neurobiologist at Columbia University and an adviser to the theatre. "Then science itself becomes a character, not just window dressing." ■

Jascha Hoffman is a writer based in New York.
e-mail: jascha@jaschahoffman.com

See http://tinyurl.com/7otvba for more details on the First Light Festival.

NEWS & VIEWS

NEUROSCIENCE

Pre-emptive blood flow

David A. Leopold

Electrical signalling among brain cells summons the local delivery of extra blood — the basis of functional brain imaging. Yet sometimes, blood is sent in anticipation of neural events that never take place.

The brain makes up only 2% of the human body mass, but because of its high energy demands it receives more than 15% of the cardiac output. When blood enters the brain, it doesn't course indiscriminately through the organ's vessels; instead, it is selectively channelled to specific regions in a need-based fashion. On page 475 of this issue, Sirotin and Das[1] show that, sometimes, blood is sent to the brain's visual cortex in the absence of any stimulus, priming the neural tissue in apparent anticipation of future events.

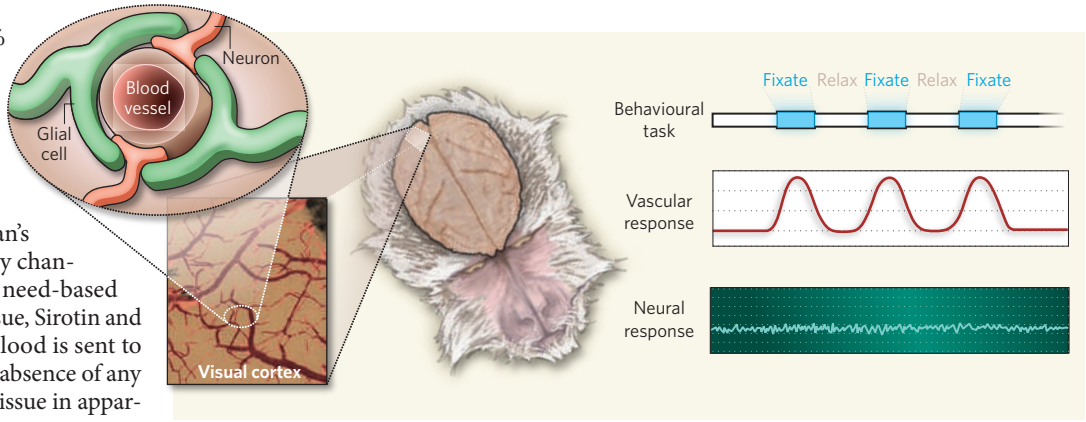


Figure 1 | Behavioural modulation of blood flow. Blood flow in the visual cortex is normally modulated in step with neural responses to visual stimuli. Sirotin and Das[1] show that the vascular response in this area of an alert monkey's brain is readily modulated by its expectation of the task, even in complete darkness and without accompanying neural modulation.

We know a lot about the physiology of the cerebral cortex, the folded sheet of densely packed grey matter that forms the outer surface of the brain. Its neurons generate electrical impulses that carry information about stimuli and events and transmit it to other neurons. Arteries, arterioles and capillaries deliver fresh blood to neurons, and supporting glial cells surround both neurons and blood vessels, regulating blood flow and performing various housekeeping roles.

Housekeeping in the brain is a challenge, because neurons undergo sudden bursts of activity that consume energy and pollute their surroundings. Consider, for example, what happens in the cortex when we first direct our gaze to a bright stimulus. In the visual cortex — the region of the cortex specialized for vision — thousands of previously quiescent neurons suddenly erupt in a cacophony of activity, each generating hundreds of electrical impulses per second. In response to the metabolic consequences of such activity, fresh blood is directed towards neurons and glia in active regions, flushing out waste products, delivering nutrients and restoring the local milieu.

Brain mapping techniques such as functional magnetic resonance imaging (fMRI) measure blood flow (haemodynamics) rather than neural activity. The accuracy with which fMRI monitors neural functioning in the human brain depends on the precise coupling between neural activity and blood flow. Although details of how neural activity triggers changes in blood supply are a topic of active debate, it is generally assumed that the two signals are tightly coupled in both space and time.

Yet biology can provide exceptions to every rule, and Sirotin and Das[1] seem to have tapped into a big one. Their study shows that cortical blood flow can depart wildly from what is expected on the basis of local neural activity. They observed this mismatch in alert rhesus monkeys by simultaneously measuring vascular and neural responses in the same region of the visual cortex. Changes in the blood supply were monitored by a sensitive video camera peering at the surface of the brain through a transparent window in the animal's skull, and local electrical responses of neurons were measured with a microelectrode. The monkeys were oblivious to all of this, focusing instead on a behavioural task that would earn them a juice reward. The task required the animals to fixate their gaze on a tiny spot on a computer monitor for several seconds at a time.

When Sirotin and Das presented the monkeys with a conventional visual stimulus under these conditions, they observed a close correspondence between neural and haemodynamic responses to the stimulus, as expected on the basis of much previous work[2]. The surprising finding came in trials without a visual stimulus, when the visual cortex should have been disengaged (Fig. 1).

Here — aside from the tiny spot directing the animal where to look — the task was carried out in complete darkness. Against all expectations, however, the haemodynamic signal continued to rise and fall. Indeed, the video of the cortical surface continued to reveal the cyclical ebb and flow of cerebral blood, which accompanied simultaneous changes in blood oxygenation and arterial diameter. By contrast, during this same trial, the neural signal fell nearly silent. Individual neurons in the same patch of cortex ceased to show any changes in their rate of generating impulses, and only the faint swell of background electrical activity indicated that cortical neurons were at all stirred by the behavioural task.

So what might be the origin of this vascular priming? One possibility is that the subtle background swells led to increased blood flow through the local release of metabolic factors. Sirotin and Das argue, however, that this interpretation is unlikely because the robust haemodynamic responses they observed did not bear a reliable relationship to the neural signals they measured. Instead, this study — perhaps more than any previous work — highlights the potential role of direct, task-related neural control of vascular tone. Both glia and blood vessels receive signals from diverse types of neuron, including signals from neurons in brain centres controlling attention and arousal[3]. The task-based modulation of the

NATURE|Vol 457|22 January 2009

vascular response might therefore represent input from a part of the brain that can anticipate probable neural activity in a specific brain region over the coming seconds, and so prime that region.

But how specific might such an anticipatory signal be? As the authors[1] found a similar modulation in other systemic markers, such as heart rate and pupil diameter, might the observed modulation reflect an overall change in the brain's blood supply? A control experiment involving an auditory task argues against this possibility. Unlike the fixation task, periodic attention to an auditory stimulus did not elicit haemodynamic modulation in the visual cortex. This control experiment, although doing little to clarify the origin of the haemodynamic modulation, shows that such priming depends on the type of sensory stimulus that the brain expects.

The mere mismatch between blood flow and neural activity per se is not a great surprise. Indeed, the precise relationship between neural activity, metabolism and blood flow has always been difficult to pin down. Early functional imaging studies revealed a quantitative discrepancy between oxygen consumption by an activated brain region following a sensory stimulus and the corresponding blood-flow response to that region[4]. Simply put, much larger amounts of oxygenated blood were delivered to an active region than were required on the basis of metabolic demands. Earlier work has also shown that the correspondence between neural activity and blood-based imaging signals is highly situation-dependent[5,6], highlighting the complex relationship between neural activity, metabolism and blood flow[7–9].

Yet the neurovascular mismatch reported by Sirotin and Das[1] is extreme. The clear and rhythmic haemodynamic modulation in the visual cortex spurred by a task performed in complete darkness is sure to raise eyebrows among the human fMRI research community. For one thing, most fMRI experiments involve the periodic presentation of sensory stimuli, and then rely on the temporal structure of the haemodynamic response for deducing local neural activity. The present study clearly demonstrates that some of the assumptions underlying such analysis — namely, that cyclical variations in blood flow reflect local, stimulus-driven events — may sometimes be incorrect. ■

David A. Leopold is in the Unit on Cognitive Neurophysiology and Imaging, Laboratory of Neuropsychology, National Institute of Mental Health, Bethesda, Maryland 20892, USA.
e-mail: leopoldd@mail.nih.gov

1. Sirotin, Y. B. & Das, A. *Nature* **457,** 475–479 (2009).
2. Logothetis, N. K. *Phil. Trans. R. Soc. Lond. B* **357,** 1003–1037 (2002).
3. Hamel, E. *J. Appl. Physiol.* **100,** 1059–1064 (2006).
4. Fox, P. T., Raichle, M. E., Mintun, M. A. & Dence, C. *Science* **241,** 462–464 (1988).
5. Maier, A. *et al. Nature Neurosci.* **11,** 1193–1200 (2008).
6. Nir, Y. *et al. Curr. Biol.* **17,** 1275–1285 (2007).
7. Devor, A. *et al. J. Neurosci.* **28,** 14347–14357 (2008).
8. Attwell, D. & Iadecola, C. *Trends Neurosci.* **25,** 621–625 (2002).
9. Logothetis, N. K. *Nature* **453,** 869–878 (2008).

ASTROPHYSICS

Galaxies in from the cold

Reinhard Genzel

Computer simulations of the cosmos suggest that cold streams of gas could underlie the unexpectedly high star-formation activity of many massive galaxies found to exist a few billion years after the Big Bang.

Recent surveys of galaxies[1] have found evidence that galaxies with masses comparable to or greater than that of the Milky Way were already present in large numbers about 3 billion years after the Big Bang. What's more, a significant fraction of these massive galaxies seem to have been gas-rich, rotating disks in which stars formed at a rate of up to 150 solar masses per year, 50 times the rate in the present-day Milky Way[2,3]. Most of these star-forming galaxies do not seem to be the aftermath of mergers of smaller systems[2], and are found to produce stars steadily over a long period of time[4] — characteristics that are at odds with the prevailing view of how galaxies form and evolve. So, if mergers are not the cause, what else could trigger the formation of stars in these massive galaxies? Building on earlier work[5–7], Dekel *et al.*[8] (page 451 of this issue) use cosmological simulations[7] to show that the galaxies might grow and form stars as a result of being fed by rapid, narrow streams of cold gas.

In the classical picture of their formation[9,10], galaxies are created when gas cools and collects at the centres of collapsing haloes of dark matter. They then evolve to form larger galaxies by merging with smaller galaxies. But the fundamental properties of atomic cooling divide evolving galaxies into two branches. Galaxies with a dark-matter mass below a critical value of about half the mass of the Milky Way (about 500 billion solar masses) would have grown rapidly through accretion of cold gas, resulting in sustained (or continuous) star formation. In contrast, massive galaxies would have grown at a much slower pace (determined by the gas cooling rate), and bursts of star formation would have occurred only when parent galaxies of comparable masses underwent intense, rapid mergers (known as major mergers).

The classical picture thus predicts that active star-forming galaxies of low mass should already have been abundant at early epochs in the history of the Universe, whereas massive galaxies should have assembled later through mergers of smaller systems. But Dekel *et al.*[8] challenge this picture using substantially improved, high-resolution hydrodynamical simulations[7] representing a large volume of the cosmos.

Perhaps the most notable aspect of Dekel and co-workers' study is the indication that, at early epochs (redshift $z \geq 2$), haloes with a mass substantially above the average for that epoch tend to form at the dense nodes of the 'cosmic web' of dark matter, which comprises long filaments of denser gas connecting these nodes. As a result, much of the gas in the filaments remains cold and flows at high speed from large distances deep into the halo, near the evolving disk in which stars will subsequently form. Under these conditions, massive galaxies above the critical mass can grow rapidly and steadily.

Dekel *et al.* find, however, that two other requirements must be met to explain the high star-formation rates of massive, early-epoch galaxies. The first is that the accretion of material must be largely gaseous, with only a small fraction of stars involved. The second is that the conversion of accreted gas into stars must be highly efficient. Because most of the gas in the streams is smooth or exists in low-mass clumps and smaller (satellite) galaxies, the first criterion means that the star-formation efficiency in the streams must be low. Qualitatively, this might be plausible if massive stars that did manage to form in the streams inject energy back into the interstellar medium through supernova explosions and stellar winds. Such 'stellar feedback' can thus disperse the surrounding gas and halt further star formation in the streams, and it would do so more effectively in lower-mass systems because of their lower gravitational binding energy. In addition, the fraction of dense molecular gas required to form stars may be lower in these lower-mass clumps, which contain fewer heavy elements and have less shielding against destructive ultraviolet radiation from hot stars and the intergalactic radiation field.

Although star formation in present-day galaxies is inefficient, the necessary high star-formation rates at early epochs might be reached if star formation consumes piled-up gas at the same rate as it is deposited. This requires that the fraction of gas ejected from the evolving massive galaxy by stellar winds and supernovae is modest, which might be at odds with observations[11]. Alternatively, or additionally, a higher efficiency of star formation or a different distribution of stellar masses may be required in the early-epoch systems compared with those in the present-day Universe. Dekel and colleagues' simulations cannot yet resolve the complex interaction between the gas inflow and the disk, which might answer many of the remaining questions. Much more detailed simulations will be required to correctly model the radiation, energy balance, dynamical structure

and star formation in the embryonic disks.

It is nevertheless tempting to conclude that the cold streams hypothesized by Dekel *et al.*[8] can explain the formation of the early-epoch massive disks. If so, what happens next in the process of galaxy formation? Gas-rich, turbulent disks are unstable, and prone to fragmentation and the formation of massive star-forming clumps of gas[12], in agreement with observations[3]. Dynamical friction then forces the clumps to spiral rapidly into the centre of the galaxy, forming a central bulge surrounded by a remnant disk[12], whose present-day relic may be the old 'thick disk' component seen in nearby galaxies. Gas accretion decreases naturally at later epochs, perhaps aided by the energy injection of massive black holes that form at the galactic centre. Disk turbulence then subsides, and a maturing thin disk can plausibly grow at redshift ≤1. Depending on whether or not a major merger occurs during this period, the end-product might be a massive elliptical or disk galaxy. The work by Dekel and co-workers may turn out to be a decisive stepping stone in elucidating the origin of these massive galaxies. ■

Reinhard Genzel is at the Max Planck Institute for Extraterrestrial Physics, 85748 Garching, Germany, and in the Department of Physics, University of California, Berkeley, USA.
e-mail: genzel@mpe.mpg.de

1. Fontana, A. *et al. Astron. Astrophys.* **459,** 745–757 (2006).
2. Shapiro, K. L. *et al. Astrophys. J.* **682,** 231–251 (2008).
3. Genzel, R. *et al. Nature* **442,** 786–789 (2006).
4. Daddi, E. *et al. Astrophys. J.* **670,** 156–172 (2007).
5. Dekel, A. & Birnboim, Y. *Mon. Not. R. Astron. Soc.* **368,** 2–20 (2006).
6. Kereš, D., Katz, N., Weinberg, D. H. & Davé, R. *Mon. Not. R. Astron. Soc.* **363,** 2–28 (2005).
7. Ocvirk, P., Pichon, C. & Teyssier, R. *Mon. Not. R. Astron. Soc.* **390,** 1326–1338 (2008).
8. Dekel, A. *et al. Nature* **457,** 451–454 (2009).
9. Rees, M. J. & Ostriker, J. P. *Mon. Not. R. Astron. Soc.* **179,** 541–559 (1977).
10. White, S. D. M. & Rees, M. J. *Mon. Not. R. Astron. Soc.* **183,** 341–358 (1978).
11. Erb, D. K. *Astrophys. J.* **674,** 151–156 (2008).
12. Bournaud, F., Elmegreen, B. G. & Elmegreen, D. M. *Astrophys. J.* **670,** 237–248 (2007).

STRUCTURAL BIOLOGY

Actin in a twist

Kenneth C. Holmes

How monomers of the cytoskeletal protein actin join to form the stable polymers crucial to muscle contraction and cellular motility has been a long-standing question. A state-of-the-art approach provides an answer.

The actin protein is abundant in all eukaryotic cells (those characterized by a membrane-bound nucleus), and is particularly prevalent in muscle, where it comprises about 20% of the total mass. Actin comes in two forms: monomeric, globular G-actin; and polymeric, filamentous F-actin. F-actin forms through polymerization of G-actin, a process that has a central role in cell motility. In non-muscle cells, for example, polymerization of F-actin induces cell movement by pushing structures such as membranes forwards[1]. Also, when one stands or walks, all of the tension causing muscle contraction is produced in muscle fibres by F-actin filaments interacting with the motor protein myosin[2,3]. Although the atomic structure of G-actin has been known for almost 20 years[4], structural details of the F-actin monomer — which is similar, but not identical, to G-actin — have remained elusive. Reporting on page 441 of this issue, Oda *et al.*[5] derive a high-resolution structure of F-actin from analysis of X-ray fibre-diffraction patterns to elucidate the transition from G- to F-actin.

The structure of G-actin has been determined independently more than 30 times. These data show that G-actin is a rather flat molecule built from two similar major domains (outer and inner) related by a pseudo-dyad, with a molecule of the nucleotide ATP, or its hydrolysed version ADP, bound between them (Fig. 1a). The terms outer and inner refer to the position of these domains in the F-actin structure. Electron-microscopy data[6] have shown F-actin to be made of two chains that turn gradually round each other to form a right-handed 'long-pitch' helix. The inner domain is closer to the axis of this helix.

By contrast, solving the structure of F-actin has been challenging. Despite a plethora of modifications to the G-actin structure — including the use of various ATP analogues, drug binding, capping proteins or crosslinking to make the G-actin look like F-actin — all of the structures derived have been essentially the same as that of the G-actin monomer, and none has yielded the secret of the G- to F-actin transition.

The most detailed data on the structure of F-actin came from X-ray diffraction of arrays of this filamentous protein oriented in a liquid-crystalline gel[7]. Diffraction from an oriented gel gives a fibre diagram — a section through the diffraction pattern of a single fibrous molecule that has been averaged by spinning it around the fibre axis (cylindrical averaging). If the molecule has periodic repeats, as in F-actin, then the fibre diagram consists of a series of lines called layer lines. The best way to interpret such a diagram is to compute it from a

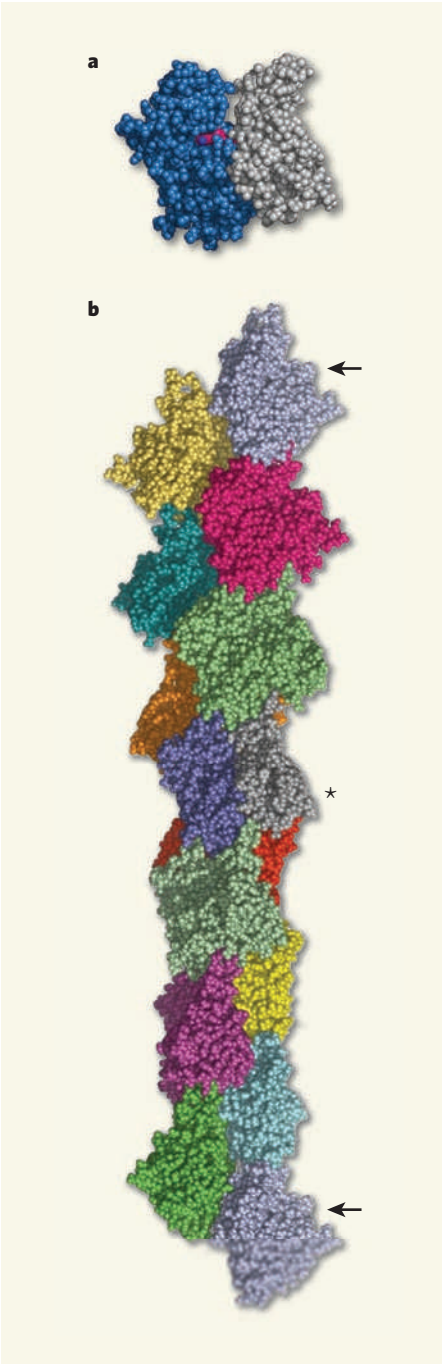


Figure 1 | Monomer versus polymer. a, G-actin monomers consist of two similar domains: outer (grey) and inner (blue). The names relate to the position of each domain in the F-actin helix — the inner domain is closer to the helix axis. The bound ADP molecule (magenta) is sandwiched between the inner and outer domains. (Data from ref. 9.) **b**, Oda *et al.*[5] provide an atomic model of the F-actin helix. This helix repeats in six left-handed turns (measuring 35.7 nanometres). Each repeat contains 13 molecules so that the first molecule of each repeat (arrowed) is in an identical orientation. Because the rotation per molecule (167°) is close to 180°, the F-actin structure appears as two long-pitch helices slowly winding around each other. Each of the 14 G-actin molecules of the helix is shown in a different colour, apart from the grey–blue molecule that is shown in two colours for comparison with **a** (asterisk).

NATURE|Vol 457|22 January 2009

50 & 100 YEARS AGO

50 YEARS AGO

With the co-operation of the West Wales Field Society, the Nature Conservancy is purchasing the island of Skomer as a National Nature Reserve. Skomer, 722 acres, is the largest of the islands off the Pembrokeshire coast ... Each year great numbers of sea-birds breed on Skomer, as well as a strong colony of the Atlantic grey seal. Probably the most notable breeding species is the Manx shearwater ... The puffin colony is, next to that of St. Kilda in the Hebrides, probably the largest in the British Isles. The island is also well known for the Skomer vole, which differs from the common bank vole in its larger size, extreme tameness and brighter colour. Skomer is being leased to the West Wales Field Society, which made an extensive field survey of the island in 1946 ... Scientific investigations on the reserve may be arranged with the Regional Officer for South Wales of the Nature Conservancy.

From *Nature* 24 January 1959.

100 YEARS AGO

Everyone who is working at radio-activity at the present time feels the need of a standard of activity in terms of which all measurements of activity can be expressed. It was suggested three years ago by Prof. H. N. McCoy ... that the activity of one square centimetre of a layer of suitable thickness of uranium oxide, U_3O_8, would furnish an excellent standard. In the December (1908) numbers of the *American Journal of Science* and of *Le Radium* Prof. McCoy gives an account of the work he has done ... to show that such a layer has all the properties required in a standard. The oxide is easily prepared, and samples prepared from three different sources gave identical results. A layer of thickness such that 0.02 gram goes to the square centimetre gives the maximum activity due to the α rays. The radiation due to the β rays is small.

From *Nature* 21 January 1909.

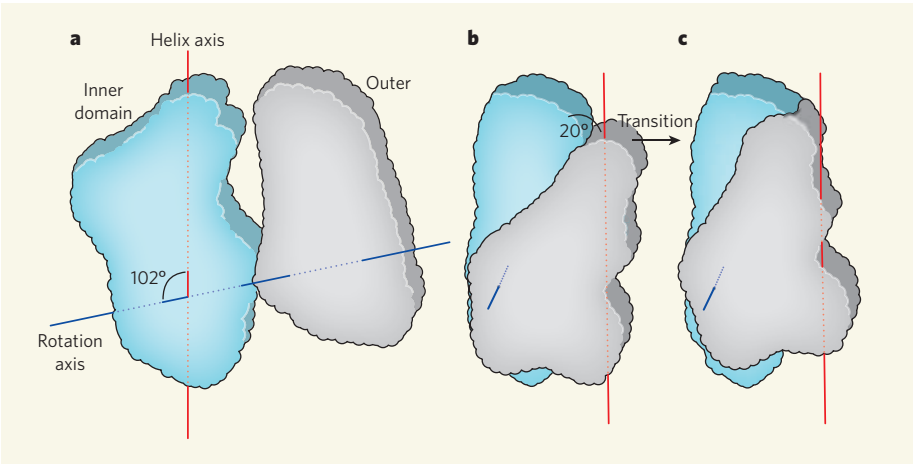


Figure 2 | The secret of G- to F-actin transition. **a**, The inner (blue) and outer (grey) domains of a single actin molecule as seen in Figure 1a. Traces of the local rotation axis and the helix axis are also shown. **b**, **c**, G-actin (**b**) and F-actin (**c**) as seen at approximately right angles compared with **a**, looking along the local rotation axis. Note that, in passing from G-actin to F-actin, the outer domain rotates by 20° with respect to the inner domain. Consequently, the F-actin structure is substantially flatter.

starting model and then use a refinement process to modify the model and arrive at a better fit with experimental data. Unfortunately, cylindrical averaging causes loss of information. Furthermore, thermal movement of the molecules in the liquid-crystalline sample causes them to become disorientated, which leads to a smearing out of the layer lines. So success depends on preparing the sample with the best possible orientation — that is, with all molecular axes as parallel to the fibre axis as possible — and investigating it at the highest possible resolution.

Oda *et al.*[5] used an intense magnetic field to improve the orientation of their sample, and a highly collimated, intense X-ray source to collect data. They started with a model made by placing the crystal structure of the G-actin molecule in the best orientation in the F-actin helix, rather like the original structure of F-actin that my colleagues and I proposed[7]. They then calculated the low-energy vibrational modes of the G-actin monomer and selected the combination of modes that best fitted the fibre diagram.

Using this improved structure, they repeated the procedure. When no further improvements could be made, the authors turned to simulated annealing. In this molecular dynamics procedure, a molecule is heated for a few picoseconds to agitate the atoms so as to sample a range of possible structures. The molecule is then slowly cooled, with the fit to the fibre diagram acting as a pseudo-force to steer the process to the correct structure[8]. So Oda *et al.* finally achieved a very good fit to the fibre diffraction pattern.

Despite the complexity of this procedure, the authors' F-actin structure[5] (Fig. 1b) is convincing in its simplicity. The transition from G- to F-actin seems to involve a 20° rotation of the outer domain with respect to the inner domain about a rotation axis roughly at right angles to the helix axis (Fig. 2). In G-actin the two domains are related by a propeller-like twist. The 20° rotation reduces this twist and flattens the molecule. Apart from this rotation and a reorientation of a flexible loop at the top of the outer domain, no other substantial change seems to occur.

Actin polymerization, essential for cell motility, is driven by ATP hydrolysis. Whereas G-actin cannot hydrolyse ATP, F-actin can. Oda *et al.* find that one effect of the 20° rotation is to bring an evolutionarily conserved glutamine residue at position 137 — which is implicated in the ATP hydrolysis mechanism — closer to the β- and γ-phosphate groups of the ATP molecule. So the rotation may be the switch for turning on the ATP-hydrolysing activity of F-actin. In addition, the flattening of monomers within F-actin substantially alters the site on this protein to which the muscle protein myosin binds — an interaction essential for muscle contraction; this could explain why myosin binds with high affinity to F-actin but not at all to G-actin. Thus, the new structure will certainly become an essential ingredient in our understanding of cell motility and muscle contraction.

Finally, the flat F-actin is very much like a bacterial analogue of actin called MreB. So Oda and colleagues' structure also lends support to the idea that actin is a bridge between eukaryotes and prokaryotic organisms such as bacteria. ■

Kenneth C. Holmes is at the Max Planck Institute for Medical Research, D69120 Heidelberg, Germany.
e-mail: holmes@mpimf-heidelberg.mpg.de

1. Dos Remedios, C. G. *et al. Physiol. Rev.* **83,** 433–473 (2003).
2. Huxley, H. E. *Science* **164,** 1356–1366 (1969).
3. Geeves, M. A. & Holmes, K. C. *Adv. Protein Chem.* **71,** 161–193 (2005).
4. Kabsch, W. *et al. Nature* **347,** 37–44 (1990).
5. Oda, T., Iwasa, M., Aihara, T., Maéda, Y. & Narita, A. *Nature* **457,** 441–445 (2009).
6. Hanson, J. & Lowy, J. *J. Mol. Biol.* **6,** 46–60 (1963).
7. Holmes, K. C. *et al. Nature* **347,** 44–49 (1990).
8. Wang, H. & Stubbs, G. *Acta Cryst.* **A49,** 504–513 (1993).
9. Otterbein, L. R., Graceffa, P. & Dominguez, R. *Science* **293,** 708–711 (2001).

CLIMATE CHANGE

Shifts in season

David J. Thomson

It's cold in winter and hot in summer. But the latest analysis illustrates the need to put observational data at the forefront of attempts to achieve a more detailed understanding of the annual temperature cycle.

It has been known for more than a century[1] that increasing the concentration of carbon dioxide in the atmosphere results in an increase in Earth's surface temperature. By contrast, it is only just over a decade since the discovery that CO_2 levels also affect the timing of the annual temperature cycle[2,3], although the details remain enigmatic.

On page 435 of this issue, Stine and colleagues[4] describe how they have updated and extended earlier studies of the annual cycle[5,6], using better spatial coverage and more recent data. They have concentrated on the temperate zones because of the dominant annual temperature cycle and, at least in the Northern Hemisphere, the reasonable spatial data coverage in those zones. As well as incorporating some technical improvements, the authors analysed the annual temperature cycle over the oceans.

The annual cycle has two distinct components, amplitude and phase. Stine and colleagues conclude that the amplitude — loosely, half the difference between summer and winter temperatures — has been decreasing over most continental areas and increasing over the oceans. The phase describes the relative timing of the periodic (seasonal) component of temperature. For the most part, the seasons occur earlier over land and later over the oceans, and Stine *et al.* estimate the terrestrial phase shift to have been 1.7 days between 1954 and 2007. This shift, and the changes in amplitude, are highly anomalous when compared with the data from between 1900 and 1953, implicating human agency as the cause.

The common perception of the timing of the seasons is more complicated because it involves both changes in the annual cycle, discussed here, and the increase in average temperature. (See Figure 1 of the Supplementary Information[4] for a graphic description of the different effects.) For example, taking the date from which the temperature usually stays above freezing as marking the start of spring, the increase in average temperature, the smaller seasonal amplitude (which implies warmer winters) and the change in phase all work in the same direction, so the observed effect is large. This is well documented in studies of bird migrations and similar phenomena[7,8], where one finds many examples of seasonal patterns shifting to an earlier date by more than a month (Fig. 1). Phase changes have also been invoked to help explain problems ranging from the theory of palaeoclimates[9] to changes in sea level[10] and even in human mortality[11].

Stine *et al.*[4] also compare their observations with the results of a suite of two dozen climate models used by the Intergovernmental Panel on Climate Change (IPCC), and the results are dismaying. Some of these models reproduce the decrease in amplitude, first shown in 1980 (ref. 12), but none predicts, or even reproduces, the change in phase. I have no personal experience with these models, so beyond a general scepticism about complicated models (perhaps best expressed by George Box's dictum, "All models are wrong but some are useful"), I cannot say why they fail. We must remember, however, that although climate models incorporate an amazing variety of effects and get many things right, they are almost certainly missing many more.

As an example, in the mid-1990s I was discussing the phase problem with members of a modelling group and learned that their model had Earth in a circular orbit with no precession. This was astonishing. First, we are trying to measure the effects of CO_2 to high accuracy — say 0.01 °C, in a system in which annual temperature extremes routinely exceed ±50 °C. Second, on an ice-age timescale, the effects of precession are immense, strong enough to be used as a clock. Third, we have known that the orbit is elliptical since Johannes Kepler in the seventeenth century, and about precession since Hipparchus (around 150 BC). The duration of the instrumental temperature record is now 1% or 2% of the 26,000-year precession cycle: when trying to measure small effects it is unwise to ignore large ones.

One should also note the contrast between the enormous computational resources used by the models and the relatively meagre effort required to analyse real data. Thus, work of the type done by Stine *et al.* is to be applauded. Ignoring the time required to assemble the data and write the programs, it probably took no more than a few seconds of computer time to show effects that were not predicted by any of the models. As Richard Feynman commented, "It doesn't matter how beautiful your theory is, it doesn't matter how smart you are. If it doesn't agree with experiment, it's wrong."

S. PANTHAKY/AFP/GETTY IMAGES

Figure 1 | Migrating greylag geese. Spring, and associated phenomena such as bird migrations, now occur earlier than at the start of the twentieth century. Stine and colleagues' data[4] quantify the seasonal shift.

NATURE|Vol 457|22 January 2009

Where do we go from here? One of many perplexing problems is the year-to-year variations in phase and amplitude in temperature data. These variations are obvious in all of the long-term temperature records and are reasonably consistent with variations in the Sun's magnetic field. We do not understand the subtle influences on climate exercised by solar effects such as the solar wind, the charged particles that flow out from the Sun. Observational evidence for such a coupling has been accumulating for decades, through both palaeoclimate data[13] and studies of the upper atmosphere[14]. However, when one has observed the Sun's acoustic oscillations in barometric pressure[15], it is possibly time to pay attention to solar observations.

The solar wind carries much more energy than is available from Edward Lorenz's butterflies, often used to 'explain' purported chaotic behaviour in climate. This raises a philosophical question, as to whether the fascination with 'chaoplexology' in climate research has resulted in a failure to take observations and statistics seriously enough. Climate may be formally chaotic, but so is Earth's orbit[16] and this has not prevented people from analysing it in exquisite detail. In my opinion, chaos, fractals, long-memory processes and their ilk should be invoked only when all of the various climate forcings have been carefully studied and all simpler explanations eliminated. We are not even close to meeting that goal.

Finally, independent of any shortcomings in the models, we must remember that the observational evidence for human influence on the climate system is overwhelming. Stine and colleagues' paper[4] adds to that evidence. If we do not stop polluting Earth's atmosphere, we may not have enough time left to develop models sophisticated enough to show what is obvious in the data now. ■

David J. Thomson is in the Department of Mathematics and Statistics, Queen's University, Kingston, Ontario K7L 3N5, Canada.
e-mail: djt@mast.queensu.ca

1. Arrhenius, S. *Phil. Mag. J. Sci.* **41,** 237–275 (1896).
2. Thomson, D. J. *Science* **268,** 59–68 (1995).
3. Thompson, R. *Int. J. Climatol.* **15,** 175–185 (1995).
4. Stine, A. R., Huybers, P. & Fung, I. Y. *Nature* **457,** 435–440 (2009).
5. Mann, M. E. & Park, J. *Geophys. Res. Lett.* **23,** 1111–1114 (1996).
6. Wallace, C. J. & Osborn, T. J. *Climate Res.* **22,** 1–11 (2002).
7. Schwartz, M. D., Ahas, R. & Aasa, A. *Global Change Biol.* **12,** 343–351 (2006).
8. Parmesan, C. *Annu. Rev. Ecol. Evol. Syst.* **37,** 637–669 (2006).
9. Jones, P. D., Briffa, K. R. & Osborn, T. J. *J. Geophys. Res.* doi:10.1029/2003JD003695 (2003).
10. Barbosa, S. M., Silva, M. E. & Fernandes, M. J. *Tellus A* **60,** 165–177 (2008).
11. McGregor, G. R., Watkin, H. A. & Cox, M. *Climate Res.* **25,** 253–263 (2004).
12. Manabe, S. & Stouffer, R. J. *J. Geophys. Res.* **85,** 5529–5554 (1980).
13. Wigley, T. M. L. *Solar Phys.* **74,** 435–471 (1981).
14. Arnold, N. *Phil. Trans. R. Soc. Lond. A* **360,** 2787–2804 (2002).
15. Thomson, D. J., Lanzerotti, L. J., Vernon, F. L., Lessard, M. R. & Smith, L. T. P. *Proc. IEEE* **95,** 1085–1132 (2007).
16. Laskar, J., Joutel, F. & Boudin, F. *Astron. Astrophys.* **270,** 522–533 (1993).

MOLECULAR BIOLOGY

Concealed enzyme coordination

Elio A. Abbondanzieri and Xiaowei Zhuang

Coordination between subunits is crucial for the proper functioning of multi-component molecular machines. A single-molecule study now allows glimpses into the mechanism used by subunits of one such machine.

Even 2,000 years ago, Aristotle had noted that the whole is more than the sum of its parts. This maxim also holds true in the cell, where enzymatic proteins frequently combine to form multimeric complexes that allow individual subunits to coordinate their activities and so perform more difficult tasks than they could alone. A prominent example of such a complex is the ring ATPases[1], in which — as their name implies — several subunits form circular complexes consisting of identical (homomeric) or non-identical (heteromeric) subunits. These enzyme complexes use energy released from the hydrolysis of ATP molecules to perform diverse cellular functions, such as DNA translocation, protein degradation and ion transport. On page 446 of this issue, Moffitt and co-workers[2] provide the first direct measurement of a single enzymatic cycle by a homomeric ring ATPase, revealing an unexpected form of coordination between the subunits.

Subunits of the various ring ATPases can coordinate their activities in different ways. For instance, the three heterodimers of the F1-ATPase act sequentially, each binding an ATP molecule and hydrolysing it in order[3] (Fig. 1a). By contrast, subunits of the L Tag helicase of simian virus 40 seem to act in concert, all six of them simultaneously binding then hydrolysing ATP molecules[4] (Fig. 1b). Subunits of the unfoldase enzyme ClpX, however, are thought to act randomly, each one hydrolysing ATP independently, with their activities probably being coordinated by the geometry of the complex[5] (Fig. 1c).

To investigate the coordination mechanism of a homomeric ring ATPase in detail, Moffitt

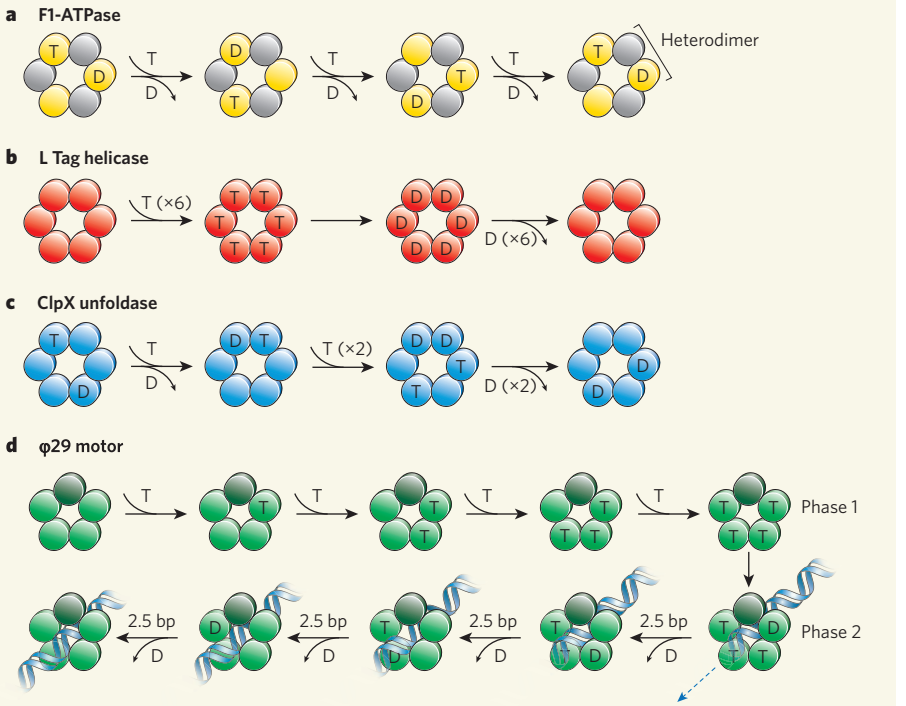


Figure 1 | Proposed coordination mechanisms of ring ATPases. a, The subunits of F1-ATPase, which exist as three heterodimers, bind to and hydrolyse ATP molecules sequentially. **b**, Subunits of the L Tag helicase, by contrast, function in concert, all simultaneously binding to ATP molecules before hydrolysing them. **c**, Subunits of the ClpX unfoldase seem to function semi-independently in a random order. **d**, Moffitt *et al.*[2] describe a newly discovered two-phase coordination mechanism for the homomeric ring ATPase of the bacteriophage ϕ29. Here, the subunits sequentially bind ATP molecules during the loading phase, and then, in a separate phase, sequentially hydrolyse ATP to translocate the DNA substrate. The exact timing of ATP hydrolysis is not known, and might not occur in conjunction with the steps. Circles indicate enzyme subunits, T denotes ATP, and D refers to its hydrolysis products.

et al.[2] used a high-precision, single-molecule assay involving dual-beam 'optical tweezers' and differential position detection[6,7] (Fig. 2). For their subject, they chose the five-subunit DNA-packaging machine of the bacteriophage virus φ29. This complex is a powerful molecular motor that can translocate the φ29 DNA into the bacteriophage's protein shell against a strong hindering force[8]. Previous circumstantial evidence suggested[9,10] that the subunits of the φ29 motor use a sequential coordination mechanism, packaging roughly 2 base pairs (bp) of DNA per ATP molecule hydrolysed. But as this putative step size was too small to be detected directly, the mechanism of subunit coordination remained largely hidden. This constraint has now been lifted, as the optical-trap assay Moffitt *et al.* report provides remarkable, sub-nanometre precision at a temporal resolution of 20 milliseconds.

The assay produced several unexpected results. The first surprise came when the authors measured DNA packaging at a low hindering force of around 8 piconewtons (pN) and found that the motor advances in steps of 10 bp, five times larger than the 2-bp step predicted by bulk biochemical assays. Given the five-subunit structure of the motor, this discrepancy could be reconciled by suggesting a concerted hydrolysis model in which the motor binds to five ATP molecules and translocates 10 bp in one step. Indeed, detailed statistical analysis of the 'dwell times' between steps, which was possible owing to the assay's high resolution, indicated that several ATP molecules must bind before the next step is taken. On closer inspection, however, the 10-bp steps seemed to be composed of smaller substeps. Although these substeps were too fast to be measured at low force, the stepping speed slowed substantially at higher forces (about 40 pN), allowing the 10-bp step to be dissected into four 2.5-bp substeps.

These striking observations raise two points. First, how are the five identical subunits of the φ29 motor coordinated to make a four-substep — rather than a more logical five-substep — advance? It could be that one of the subunits has a role, other than directly translocating DNA, that is not readily detectable. Alternatively, one of the subunits might be truly passive. The second point is that the non-integer step size of 2.5 bp seems to be intuitively at odds with our knowledge of how motor proteins move on periodic tracks. But there is no a priori rule that each subunit of an enzyme complex must make specific and identical contacts with its substrate. The interaction between the φ29 motor and DNA may be somewhat promiscuous, and the step size could be dictated by a conformational change within the enzyme, rather than by the periodicity of the substrate's structure.

Moffitt and colleagues' data[2] therefore challenge the conventional models of how DNA-processing enzymes interact with their substrates. What's more, their results suggest an exquisite coordination between the subunits of the φ29 motor ATPase that combines features of both sequential and concerted coordination mechanisms: before any DNA translocation occurs, all ATP molecules must bind during the long dwell time preceding the 10-bp advance; once loaded with ATP, the motor hydrolyses ATP and/or releases its hydrolysis products (ADP and inorganic phosphate) sequentially to produce the four successive 2.5-bp substeps (Fig. 1d).

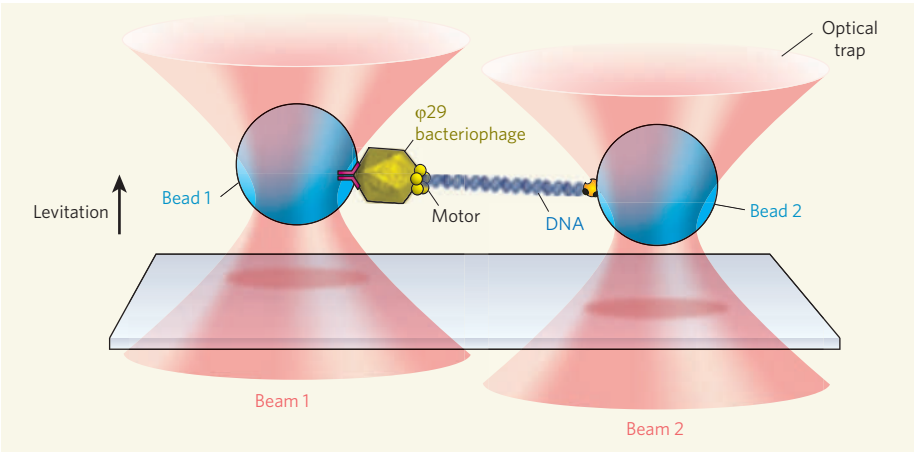


Figure 2 | Analysis of DNA translocation by the bacteriophage ϕ29 motor. Moffitt *et al.*[2] used a dual-beam optical-tweezer assay to study the packaging of DNA by the motor ATPase of ϕ29. They attached the ϕ29 protein shell with the packaging motor to a bead and the distal end of the DNA to be packaged to a second bead. The beads were levitated in separate optical traps, and the amount of DNA packaged was determined by measuring the relative distance between the beads, allowing sub-nanometre measurement precision with millisecond temporal resolution.

The cooperative binding of ATP, as implied by this model, would seem to be at odds with the previously measured[10] 'Michaelis–Menten' — rather than the expected sigmoidal — form of dependence of the DNA-packaging velocity on ATP concentration. Moffitt *et al.* resolve this apparent paradox with an elegant kinetic analysis, which shows that the expected simple Michaelis–Menten behaviour could be recovered if the binding of each ATP molecule is ordered in time and is separated by an irreversible event. This finding also indicates that each subunit must bind to an ATP molecule sequentially. Furthermore, temporal segregation between a sequential ATP-binding phase and a sequential motor-stepping phase would require coordination both between neighbouring subunits and across the entire complex.

The two-phase coordination mechanism is complicated, but could be tested further. In previous single-molecule studies of the F1-ATPase, fluorescent ATP was used to establish that, at any given time, at least two binding sites on this enzyme complex are occupied by ATP or ADP, and that the steps taken by the complex are linked to ATP binding and the release of its hydrolysis products[11]. A similar approach could be used to capture the sequential loading of ATP and unloading of its hydrolysis products in the φ29 motor, and to find the precise placement of the mechanical step in the ATP hydrolysis cycle.

Moffitt and co-workers' two-phase model may prove to be a general mechanism applicable to other ring ATPases, thus explaining why some of these enzymes have a seemingly sequential coordination mechanism of action but also bind to several ATP molecules at once[12]. Single-molecule approaches could help to unravel the coordination mechanism of these systems. The optical-tweezer assay, for example, is known to be remarkably adaptable for examining helicase, polymerase and translocase enzymes, which operate on nucleic-acid substrates. Even folded proteins have been pulled apart with optical tweezers[13], opening up the possibility of also studying unfoldase, protease or chaperone enzymes, which operate on amino-acid substrates. It will be exciting to learn whether such a diverse group of enzymes share common strategies for coordination, or if other, as-yet-unidentified forms of subunit organization exist. ■

Elio A. Abbondanzieri and Xiaowei Zhuang are at the Howard Hughes Medical Institute and Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138, USA.
e-mail: zhuang@chemistry.harvard.edu

1. Erzberger, J. P. & Berger, J. M. *Annu. Rev. Biophys. Biomol. Struct.* **35,** 93–114 (2006).
2. Moffitt, J. R. *et al. Nature* **457,** 446–450 (2009).
3. Adachi, K. *et al. Cell* **130,** 309–321 (2007).
4. Gai, D., Zhao, R., Li, D., Finkielstein, C. V. & Chen, X. S. *Cell* **119,** 47–60 (2004).
5. Martin, A., Baker, T. A. & Sauer, R. T. *Nature* **437,** 1115–1120 (2005).
6. Abbondanzieri, E. A., Greenleaf, W. J., Shaevitz, J. W., Landick, R. & Block, S. M. *Nature* **438,** 460–465 (2005).
7. Moffitt, J. R., Chemla, Y. R., Izhaky, D. & Bustamante, C. *Proc. Natl Acad. Sci. USA* **103,** 9006–9011 (2006).
8. Smith, D. E. *et al. Nature* **413,** 748–752 (2001).
9. Guo, P., Peterson, C. & Anderson, D. *J. Mol. Biol.* **197,** 229–236 (1987).
10. Chemla, Y. R. *et al. Cell* **122,** 683–692 (2005).
11. Nishizaka, T. *et al. Nature Struct. Mol. Biol.* **11,** 142–148 (2004).
12. Adelman, J. L. *et al. Mol. Cell* **22,** 611–621 (2006).
13. Cecconi, C., Shank, E. A., Bustamante, C. & Marqusee, S. *Science* **309,** 2057–2060 (2005).

OBITUARY

Daniel Carleton Gajdusek (1923–2008)

The most outlandish and peripatetic of microbe hunters.

Daniel Carleton Gajdusek, who won the 1976 Nobel Prize in Physiology or Medicine for his discovery of transmissible dementias, died on 12 December 2008 in a hotel in Tromsø, Norway. This isolated spot above the Arctic Circle was the winter refuge for American-born Gajdusek during the last 10 years of his life. He spent his summers mainly in Amsterdam. This migratory pattern, and his choice to spend winter in one of the darkest places on Earth, typified the eccentricity of the man.

Gajdusek was born in 1923 in Yonkers, New York, where his father had a butcher's shop. As an eight year old, he already seemed to know his destiny. He told me once that he inscribed the names of all the scientists in Paul de Kruif's book *Microbe Hunters* — which included giants such as Robert Koch and Louis Pasteur — on the staircase to his attic chemistry lab, leaving the last step blank for himself.

Gajdusek's mind was continuously attracted to the mysterious and exceptional, his rationale being that, to contribute to knowledge, you must find unexplained phenomena and observe them first-hand. From the early 1950s onwards, having trained as a research virologist, he recorded his scientific endeavours on film so that he could share his experiences as directly as possible with everybody.

In 1954, Gajdusek shot a documentary entitled *Rabies in Man*, which followed experiments at the Pasteur Institute of Iran in Tehran. The institute's director, Marcel Baltazard, had recently shown that almost a third of people who had suffered a rabid-dog bite to the head could not be saved by rabies vaccine. Baltazard considered this result disastrous. Gajdusek suggested that he should test anti-rabies antibodies (prepared from rabbit serum by Herald Cox in New York) in combination with the vaccine.

Baltazard agreed to this, and in August 1954 he began using the combination therapy in 18 patients who had sustained head wounds from rabid wolves. Gajdusek's documentary meticulously followed their progress during treatment. The study showed convincingly that addition of rabies antibodies to the vaccine can completely protect people from disease or death after exposure to rabies virus. This regimen has been the gold standard of care for the disease ever since.

Gajdusek's views on microbiology were shaped by the training in physical chemistry that he received from Linus Pauling, and by schooling in cell biology and virology from John Enders — both Nobel laureates. From 1955 to 1957, he also worked in

AP

Melbourne, Australia, with Frank Macfarlane Burnet, who received a Nobel prize in 1960 for his work on the recognition of 'self' by the immune system. His experiences in this vibrant field taught him to expect the unexpected, and prepared him for his greatest discovery.

In 1957, Gajdusek travelled to New Guinea upon hearing that Vincent Zigas, a district medical officer, had stumbled across a mysterious illness — kuru — in the Fore tribe of the Highlands. This turned out to be a neurological disease affecting women and children. It progressed swiftly from an initially unsteady gait to tremors and speech disorders, leading within months to complete incapacitation, and invariably to death. Gajdusek suspected from the beginning that the disease was caused by a form of ritual cannibalism in which only women and children participated.

In 1961, Gajdusek convinced Clarence Joseph Gibbs Jr, a specialist in insect-borne viruses, to lead a series of experiments designed to establish the concept of 'transmissible spongiform encephalopathies' (TSEs). Besides kuru, TSEs include neurodegenerative illnesses such as Creutzfeldt–Jakob disease (CJD) and 'mad cow' disease. Gajdusek and Gibbs reported the successful transmission of kuru to chimpanzees in 1966, of CJD to chimpanzees in 1968, and of scrapie (the sheep variant of TSE) to monkeys in 1972. In later years, CJD was categorized with Alzheimer's disease as an amyloidosis (diseases characterized by the deposition of insoluble proteins), but Gajdusek and co-workers showed in 1980 that Alzheimer's disease, unlike CJD, was not transmissible.

Gajdusek's work revealed the existence of a new kind of infectious agent, one that did not need a nucleic acid to replicate. Now called prions, these agents are misfolded proteins that can induce misfolding in other proteins. The initiation of misfolding falls in the twilight zone between normal and abnormal protein production, and is still not understood. The conspicuously unconventional Gajdusek had thus found a suitably eccentric infectious agent. The discovery won him a Nobel prize, and earned him the right to add his name to the staircase of the world's great microbe hunters.

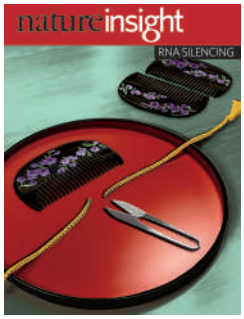
Eccentricity was the source of Gajdusek's genius as a scientist, and of his notoriety late in life. In 1997, he was imprisoned on a child molestation charge involving one of the more than 50 Micronesian and Melanesian children he had adopted and brought to the United States. On his release in 1998 he moved to Europe, which he regarded as less puritanical than his home country.

Throughout his life, Gajdusek was a fervent reader of the world's literature and a prolific writer. He believed in a life of learning and in accurate documentation and reflection. An example of his acuity was his letter writing. When you wrote to him, you got your own letter back with an answer to each sentence scribbled between the lines.

Around two months before his death, I had dinner with him at the Academic Club of the University of Amsterdam, about a minute's walk from his university lodgings. It took us at least 20 minutes to get there, stopping every minute to rest because of his failing heart. This cumbersome trip did not stop him from talking all the way about his most recent interest: the physical evidence in the brain revealing a person's reading ability and the development of that skill. He kept yelling at me, while gasping for air, that before the age of six a child could achieve native fluency in at least six, if not ten, languages, if properly exposed. He must have had his own youth in mind: Gajdusek could read at least ten languages.

Gajdusek will be remembered for both his scientific contributions and his overwhelming presence. As Richard Rhodes observed in his book on TSEs, *Deadly Feasts*, Gajdusek was "A compulsive talker who spills ideas nonstop for hours — good talk, often brilliant talk and consummate story-telling, but more than some listeners can bear".

Jaap Goudsmit
Jaap Goudsmit is in the Research and Development Department of Crucell Holland, PO Box 2048, Leiden, 2301 CA, the Netherlands, and in the Academic Medical Center of the University of Amsterdam.
e-mail: j.goudsmit@crucell.com



RNA SILENCING

Cover illustration
In RNA silencing, one strand of a small duplex RNA (combs) enters a silencing complex (platter) that contains a catalytic Argonaute (pincers) to cleave a target RNA (cord). (Courtesy of M. Inudo and Y. Tomari. Artwork by N. Spencer)

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel: +44 (0) 20 7833 4000
e-mail: nature@nature.com

nature publishing group

When *Nature* published the first Insight on RNA interference (RNAi), in September 2004, it was clear that RNAi was going to have a broad impact on biology, even though only six years had passed since the seminal paper by Andrew Fire, Craig Mello and colleagues was published.

But who would have imagined how far we would come in the next four years in terms of understanding and exploiting this fundamental system of gene regulation? There is now a much clearer picture of how the small non-coding RNAs involved in this type of regulation are generated, drawn from the static images provided by crystallographic studies, together with the kinetic and mechanistic details gleaned through biochemical assays. From large-scale efforts to map how gene expression is affected by just one class of these small RNAs, microRNAs, it is easy to reach the conclusion that when studying any biological process, researchers must consider how it is regulated by small RNAs. Relationships between small RNAs and development are also being uncovered almost daily. And nimble biotechnology firms have, with breathtaking speed, aggressively translated this knowledge into therapeutic candidates.

It was no surprise that the researchers who opened this Pandora's box were awarded the Nobel Prize in Physiology or Medicine in 2006. As Göran Hansson stated in his presentation speech for the award, RNAi "has added a new dimension to our understanding of life and provided new tools for medicine". However, the story is far from complete even now. With advances in sequencing technology, for example, more classes of small RNA are being identified, and their functions are likely to continue to entice and surprise us.

With these reviews, we hope to convey some of the excitement driving this rapidly evolving field forward. We are pleased to acknowledge the financial support of Alnylam Pharmaceuticals and Roche in producing this Insight. As always, *Nature* carries sole responsibility for editorial content and peer review.

Angela K. Eggleston, Senior Editor

REVIEWS

nature insight

NATURE|Vol 457|22 January 2009|doi:10.1038/nature07754

On the road to reading the RNA-interference code

Haruhiko Siomi[1] & Mikiko C. Siomi[1,2]

The finding that sequence-specific gene silencing occurs in response to the presence of double-stranded RNAs has had an enormous impact on biology, uncovering an unsuspected level of regulation of gene expression. This process, known as RNA interference (RNAi) or RNA silencing, involves small non-coding RNAs, which associate with nuclease-containing regulatory complexes and then pair with complementary messenger RNA targets, thereby preventing the expression of these mRNAs. Remarkable progress has been made towards understanding the underlying mechanisms of RNAi, raising the prospect of deciphering the 'RNAi code' that, like transcription factors, allows the fine-tuning and networking of complex suites of gene activity, thereby specifying cellular physiology and development.

The discovery of RNA interference (RNAi)[1] heralded a revolution in RNA biology. Researchers uncovered 'hidden' layers of regulation of gene expression, in which many previously unidentified families of small RNAs (consisting of ~20–30 nucleotides) mediate gene silencing. A diverse set of gene-regulatory mechanisms were found to use key steps in the RNAi process, including mechanisms that silence endogenous genes and mechanisms that restrain the expression of parasitic and pathogenic invaders such as transposons and viruses[2–5].

The basic RNAi process can be divided into three steps[6,7]. First, a long double-stranded RNA (dsRNA) that is expressed in, or introduced into, the cell (for example, as a result of the base-pairing of sense and antisense transcripts or the formation of stem–loop structures) is processed into small RNA duplexes by a ribonuclease III (RNaseIII) enzyme known as Dicer. Second, these duplexes are unwound, and one strand is preferentially loaded into a protein complex known as the RNA-induced silencing complex (RISC). Third, this complex effectively searches the transcriptome and finds potential target RNAs. The loaded single-stranded RNA (ssRNA), called the guide strand, then directs an endonuclease that is present in the RISC (sometimes called the 'slicer' and now known to be an Argonaute protein[8–11]) to cleave messenger RNAs that contain sequence homologous to the ssRNA, over many rounds. In this way, the guide strand determines the sequence specificity of the RNAi response.

In different organisms, the RNAi pathways comprise different proteins and mechanisms, but they operate by strikingly convergent strategies. In all organisms that have been studied, RNAi involves two main components: small RNAs, which determine the specificity of the response; and Argonaute proteins, which carry out the repression. Depending on both the nature of the Argonaute in the RISC and the degree of complementarity between the small RNA and the target sequence in the mRNA, the association of the RISC with target mRNAs has been shown to have different outcomes: it can control protein synthesis and mRNA stability, maintain genome integrity or produce a specific set of small RNAs[8,12]. Analyses of the biogenesis of small RNAs and their targeting mechanisms have benefited from the advent of high-throughput sequencing technologies and sophisticated bioinformatics[13]. The picture emerging from these studies is that RNAi systems in different organisms have been refined in many ways, and such modifications include built-in molecular 'rulers' that define the size of small RNAs, structures that determine which strand of a small RNA is selected, mechanisms that direct further rounds of small RNA amplification, or safeguards against off-target (unrestricted and unrelated) silencing.

Another emerging finding in the field is that the activity of RNAi pathways is subject to intense regulation at various levels, from the level of biogenesis of small RNAs to the silencing mode of the RISC. In this Review, we describe the biogenesis of the guide strand of small RNAs and the formation and actions of the RISC, and we discuss the current understanding of the molecular mechanisms of RNAi in the light of recent insights into how silencing pathways are specified and regulated.

Biogenesis of small RNAs

A hallmark of RNAi is that short (~20–30 nucleotide) dsRNAs — known as small RNAs — are generated by the activity of RNaseIII enzymes (either Dicer alone or Drosha and Dicer). Two main categories of small RNAs have been defined on the basis of their precursors. The cleavage of exogenous long dsRNA precursors in response to viral infection or after artificial introduction generates short interfering RNAs (siRNAs), whereas the processing of genome-encoded stem–loop structures generates microRNAs (miRNAs). Using high-throughput sequencing technology, several new classes of endogenous small RNA species have recently been uncovered, and these include PIWI-interacting RNAs (piRNAs) and endogenous siRNAs (endo-siRNAs or esiRNAs).

A common feature of all of these small RNAs is that they are loaded onto Argonaute proteins to effect their targeting function (discussed further in the section 'Loading and sorting of small RNAs by the RISC'). An overview of the generation of small RNAs is presented in Fig. 1.

siRNA biogenesis

Dicer (Table 1) processes long RNA duplexes and generates siRNAs. These small RNAs are ~21–25-nucleotide duplexes with a phosphate group at both 5′ ends, and hydroxyl groups and two-nucleotide overhangs at both 3′ ends, all hallmarks of RNaseIII-mediated cleavage. The Dicer protein contains a PAZ domain, which binds to the 3′ end of an siRNA, and two RNaseIII domains, which have the catalytic activity. It functions as a monomer[14], but the RNaseIII domains associate with each other to

[1]Department of Molecular Biology, Keio University School of Medicine, 35 Shinanomachi, Shinjuku-ku, Tokyo 160-8582, Japan. [2]Japan Science and Technology Agency (JST), Core Research for Evolutional Science and Technology (CREST), 4-1-8 Hon-chou, Kawaguchi, Saitama 332-0012, Japan.

form an 'internal dimer' (see page 405). The distance between the PAZ domain and the two RNaseIII domains is the length spanned by 25 base pairs (bp) of RNA[15]. Thus, Dicer itself is a molecular ruler.

miRNA biogenesis

Similarly, miRNAs are short (~21–25-nucleotide) RNA molecules[16]; however, their biogenesis differs markedly from that of siRNAs. The primary precursors of miRNAs (pri-miRNAs) are encoded in the genome, and the relevant genomic regions are mostly transcribed by RNA polymerase II (ref. 17). The pri-miRNAs contain stem–loop structures that harbour the miRNA in the 5′ or 3′ half of the stem. During miRNA production in plants, one type of RNaseIII, Dicer-like protein 1 (DCL1), generates the miRNA–miRNA* duplex in the nucleus (miRNA* being the sequence in the stem–loop that pairs with the miRNA, equivalent to the passenger strand of siRNA duplexes; discussed later). By contrast, in animals, miRNAs are derived in a two-step process, in which the nuclear-localized RNaseIII Drosha defines one end of the miRNA–miRNA* duplex and releases a precursor miRNA (pre-miRNA) of ~65–70 nucleotides. The pre-miRNA hairpin is then exported to the cytoplasm, where Dicer completes the processing.

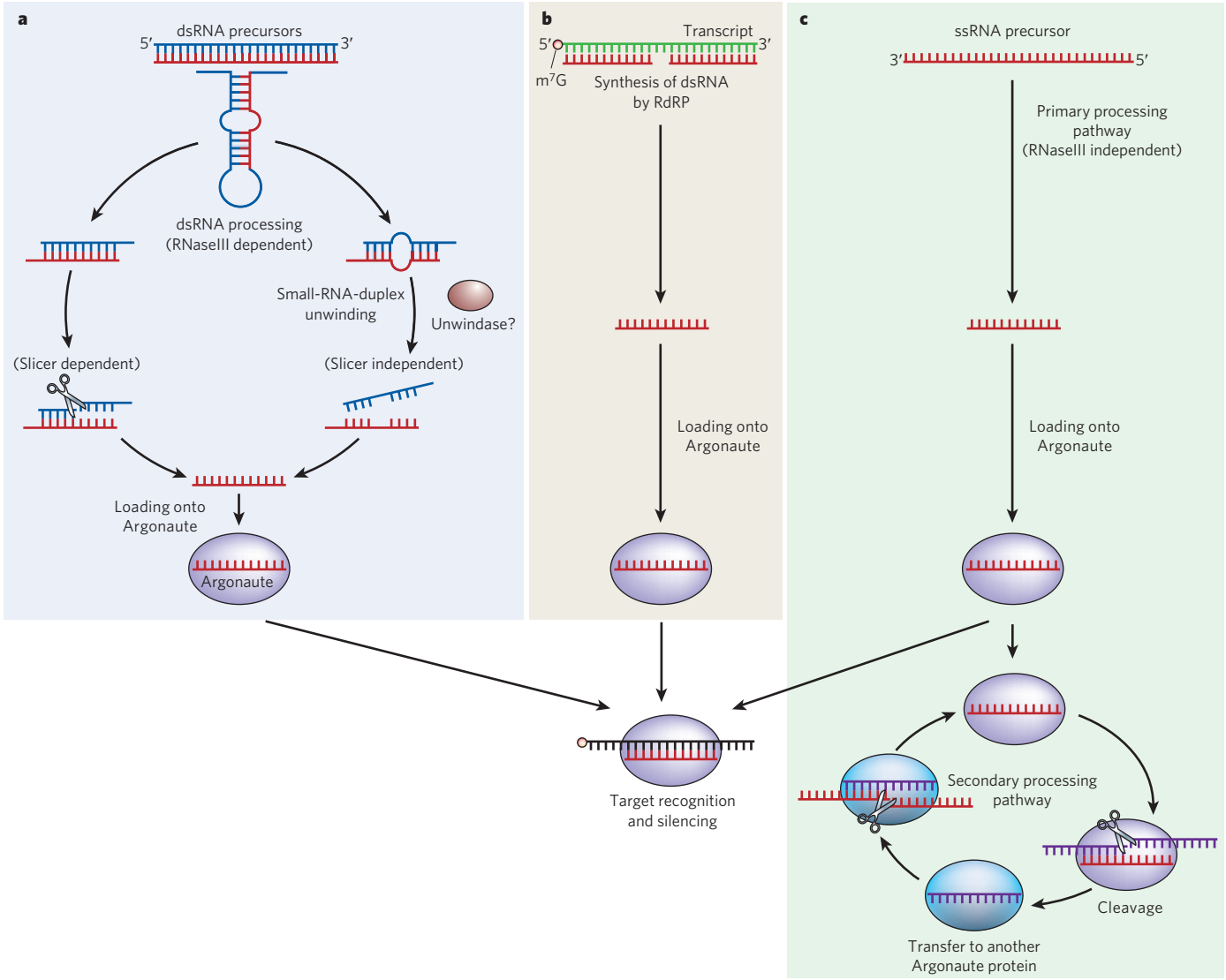


Figure 1 | Small RNA production and RNA silencing. a, Natural transcripts that form dsRNAs and hairpin-shaped structures can be sources of small RNAs. These precursors are processed by an RNaseIII enzyme (such as Drosha or Dicer), yielding small RNA duplexes. Duplexes with a perfect match (left pathway) are further processed by an enzyme with slicer activity (an Argonaute protein) into single-stranded small RNAs. By contrast, small RNA duplexes with a mismatch or bulge in the centre (right pathway) are not substrates for the slicer and thus become single-stranded in a cleavage-independent manner. The identity of the protein that carries out this unwinding is unknown. Single-stranded small RNAs are then loaded onto Argonaute proteins. The particular strand that is selected (sense or antisense) depends on thermodynamic stability. The loaded Argonaute proteins are guided to target mRNAs containing complementary sequence, and the expression of the corresponding genes is silenced. The mode of action of this silencing — whether the mRNA is cleaved or whether translation is just repressed — largely depends on the degree of complementarity between the target mRNAs and the Argonaute-associated small RNAs. **b**, Some small RNAs found in *Caenorhabditis elegans* and plants are known to be produced in an RNA-dependent RNA polymerase (RdRP)-dependent manner. Natural transcripts (often aberrant RNAs) can be substrates for this type of small RNA synthesis. This does not occur in organisms that lack RdRP activity, such as mammals and *Drosophila melanogaster*. Single-stranded small RNAs generated in this way can then be loaded onto Argonaute proteins and silence gene expression. **c**, The PIWI subfamily of Argonaute proteins, which are germline specific, are loaded with piRNAs. These complexes function to silence transposons. Single-stranded precursors give rise to piRNAs, through a mechanism called the primary processing pathway. The proteins required for this pathway are unknown. The silencing of transposons by PIWI proteins simultaneously amplifies piRNAs in germ cells. This pathway is known as the secondary processing pathway (or the ping-pong amplification loop) and is conserved in a variety of organisms, including mice and zebrafish. In this pathway, the slicer activity of the PIWI proteins reciprocally forms the 5′ ends of piRNAs by cleaving transposon transcripts (piRNA precursors). Proteins required to form the 3′ end of piRNAs remain unidentified.

Table 1 | Key proteins in RNA silencing in various organisms

Protein	Yeast (*Schizosaccharomyces pombe*)	Plant (*Arabidopsis thaliana*)	Nematode (*Caenorhabditis elegans*)	Fruitfly (*Drosophila melanogaster*)	Mammal	
					Mouse	Human
RNaseIII	Dcr1	DCL1	DCR-1	DCR-1	DICER1	DICER1
		DCL2	DRSH-1	DCR-2	DROSHA	DROSHA
		DCL3		DROSHA		
		DCL4				
Argonaute: AGO subfamily	Ago1	AGO1	ALG-1	AGO1	AGO1	AGO1
		AGO2	ALG-2	AGO2	AGO2	AGO2
		AGO4	3 others		AGO3	AGO3
		AGO5			AGO4	AGO4
		AGO6			AGO5 (possibly a pseudogene)	
		AGO7 (ZIPPY)				
		AGO10 (ZLL, PNH)				
		3 others				
Argonaute: PIWI subfamily	None	None	ERGO-1	AGO3	MILI (PIWIL2)	HILI (PIWIL2)
			PRG-1	PIWI	MIWI (PIWIL1)	HIWI (PIWIL1)
			PRG-2	AUB	MIWI2 (PIWIL4)	HIWI2 (PIWIL4)
						PIWIL3 (HIWI3)
Argonaute: WAGO subfamily	None	None	RDE-1	None	None	None
			SAGO-1			
			SAGO-2			
			PPW-1			
			PPW-2			
			CSR-1			
			NRDE-3			
			11 others			
Double-stranded-RNA-binding domain (dsRBD)-containing cofactor of RNaseIII	None	HYL1	PASH-1	PASHA	DGCR8	DGCR8
			RDE-4	R2D2	TRBP (TARBP2)	TRBP (TARBP2)
				LOQS	PACT (PRKRA)	PACT (PRKRA)
RNA-dependent RNA polymerase (RdRP)	Rdp1	RDR1	EGO-1	None	None	None
		RDR2 (SMD1)	RRF-1			
		RDR6 (SDE1, SGS2)	RRF-3			
		3 others	1 other			

Molecules that belong to these categories but have unknown functions are not listed but are indicated as 'others'. Common synonyms are indicated in parentheses. Data were taken from refs 8, 12, 29, 50 and 98.

Drosha is present in a large complex, known as the microprocessor complex, which functions like a molecular ruler to determine the cleavage site in the pri-miRNAs[18,19]. In this complex, Drosha interacts with its cofactor, known as DGCR8 or Pasha (depending on the species), which also binds to dsRNA (through its dsRNA-binding domain; dsRBD)[20,21]. A typical metazoan pri-miRNA consists of a 33-bp stem, a terminal loop and ssRNA flanking segments. The flanking segments are crucial for binding to DGCR8, and the 33-bp stem is also required for efficient binding. Drosha can interact transiently with the stem of this 'pre-cleavage' complex, and the processing centre of the enzyme, located at ~11 bp from the ssRNA–dsRNA junction, makes a staggered pair of breaks in the RNA to create the ~65–70-nucleotide pre-miRNA. Thus, DGCR8 might function as the molecular anchor that measures the distance from the ssRNA–dsRNA junction. It is possible that the microprocessor complex could recognize the terminal loop as ssRNA and bind to the stem–loop structure in the opposite orientation. In this case, abortive cleavage can occur at an alternative site ~11 bp from the terminal loop. However, most pri-miRNAs contain internal bulges or weakly paired bases ~11 bp from the terminal loop that mitigate processing from this direction[18].

Although many of the sequences encoding miRNAs are located within introns, clusters encoding miRNAs that are processed directly by the spliceosome, instead of Drosha, were recently identified[22,23]. The 3′ end of the stem–loop precursor of these intronic miRNAs (known as mirtrons) coincides with the 3′ splice site of a small annotated intron and is cleaved in the same splicing pathway as pre-mRNA in the nucleus instead of by Drosha. Subsequently, the mirtron precursors, which are released by the spliceosome in the shape of a lariat (lasso), are linearized by a de-branching enzyme. They then enter the miRNA-processing pathway directly (by mimicking the structural features of pre-miRNA hairpins) and are therefore exported to the cytoplasm and processed by a Dicer protein, bypassing Drosha-mediated cleavage.

The imprecision of Drosha or Dicer cleavage could result in the production of a set of miRNA–miRNA* duplexes with a variety of 5′ and 3′ ends. Most miRNAs in animals form imperfect hybrids with sequences in the target mRNA, with most of the pairing specificity being provided by the 5′-proximal region of the miRNA (that is, positions 2–8; also known as the seed region)[24,25]. Imprecise cleavage either alters the seed sequence or inverts the relative stabilities of the 5′ and 3′ ends of the duplex (see the section 'Loading and sorting of small RNAs by the RISC'). The results of recent deep-sequencing studies of small RNAs, however, indicate that human cells might take advantage of such imprecise cleavage, because the generation of a diverse set of miRNAs from a single precursor could be a way of broadening the network of factors and processes that are regulated by miRNAs[26–28].

RNaseIII-independent pathways of small RNA biogenesis

In some systems, small RNAs do not seem to be produced in response to dsRNA, but silencing signals are still amplified. Because these small RNAs do not arise from dsRNA precursors, RNaseIII enzymes cannot be involved in their generation. These findings therefore call into question the definition of RNAi. In this subsection, we describe the known RNaseIII-independent pathways of small RNA production, including those that generate piRNAs, 21U-RNAs, and secondary siRNAs in *Caenorhabditis elegans*.

The small RNAs known as piRNAs were named for their ability to bind to a group of Argonaute proteins known as PIWI proteins. As noted earlier, members of the Argonaute family bind directly to small guide RNAs and lie at the core of all known RISCs[8]. Argonaute proteins consist of a variable amino-terminal domain and three conserved domains (the PAZ, middle (MID) and PIWI domains)[8,29,30]. The 3′ end of a small RNA interacts with the PAZ domain, whereas the phosphate group at the 5′ end of small RNAs binds to a cleft bridging the MID domain and the PIWI domain[29,30] (see page 405). The PIWI domain has an RNaseH-like folded structure[10] and slicer activity (although some Argonaute proteins seem to have no slicer activity). There are three phylogenetic groups of Argonaute proteins[29]: the AGO subfamily (or AGO clade), named after the founding member *Arabidopsis thaliana* ARGONAUTE 1 (AGO1); the PIWI subfamily, named after *D. melanogaster* PIWI (*P*-element-induced wimpy testis); and the WAGO (worm-specific Argonaute) subfamily of *C. elegans*-specific proteins. PIWI-subfamily proteins bind to piRNAs[31–37] (Table 1). These small RNAs have been found only in germ cells, and they are important for germline development and suppress transposon activity in the germline cells of mammals, fish and *D. melanogaster*. They are ~24–31 nucleotides (slightly longer than miRNAs), usually have a uridine at the 5′ end and carry a 5′ monophosphate. Unlike mammalian miRNAs, but similarly to plant miRNAs, piRNAs have a 2′-*O*-methyl (2′-*O*-Me) modification on the nucleotide at the 3′ end, a modification that is carried out by a HEN1-like methyltransferase[38–42]. If Dicer is mutated, the production of piRNAs is not affected, indicating that their biogenesis is distinct from that of miRNAs and siRNAs and does not involve dsRNA precursors[31,42].

The sequencing of small RNAs associated with *D. melanogaster* PIWI-subfamily proteins (PIWI, Aubergine (AUB) and AGO3)[43,44] showed that piRNAs associated with AUB and PIWI are derived mainly from the antisense strand of retrotransposons, whereas AGO3-associated piRNAs arise mainly from the sense strand. AUB- and PIWI-associated piRNAs show a strong preference for uridine at their 5′ ends, whereas AGO3-associated piRNAs show a preference for adenosine at nucleotide 10. The first ten nucleotides of AUB-associated piRNAs can be complementary to the first ten nucleotides of AGO3-associated piRNAs. In addition, PIWI-subfamily proteins have slicer activity that allows them to cleave an RNA substrate opposite position 10 of their bound piRNA[32,44]. These observations suggest that piRNAs have a self-amplifying loop (Fig. 1), in which sense piRNAs associated with AGO3 cleave long antisense transcripts and guide the formation of the 5′ end of antisense piRNAs bound to AUB or PIWI, and vice versa. Thus, in this amplification loop, which is called the ping-pong cycle[43], transposons are both a source of piRNAs and a target of piRNA-mediated silencing. After the resultant cleavage products have been loaded onto another member of the PIWI subfamily, further (as yet unidentified) nuclease activity generates the 3′ end of the piRNA, with the specific size of the piRNA determined by the footprint of the PIWI-subfamily protein on the RNA, a step that seems to precede 2′-*O*-Me modification[38]. In each PIWI-subfamily protein, the PAZ domain might be positioned at a distance from the MID domain that corresponds to the length of each piRNA. Thus, the PAZ domain might function as part of a molecular ruler for processing piRNAs of a defined size. Signatures of this amplification cycle are also apparent in zebrafish (*Danio rerio*) germ cells and in mammalian germ cells before the pachytene stage of meiosis during spermatogenesis[42,45].

PIWI-subfamily proteins and, presumably, their associated piRNAs are loaded into embryos from the ova[8], implying that the piRNAs that initiate an amplification cycle of piRNA biogenesis (which generates secondary piRNAs) could be supplied by germline transmission. But several findings indicate that there must be mechanisms of piRNA biogenesis other than amplification induced by maternal piRNAs. First, the amplification cycle in *D. melanogaster* engages mainly AGO3 and AUB[43,44], but piRNAs are still loaded onto PIWI, which is spatially separated from these proteins at the subcellular and cell-type levels[32,43,44]. Second, piRNAs derived from a particular piRNA cluster in the genome (the *flamenco* locus) associate almost exclusively with PIWI[43]. These findings indicate that *flamenco*-derived piRNAs are produced by a pathway independent of the amplification loop. Whether such a piRNA-biogenesis pathway exists remains to be determined.

What at first seemed to be another type of small RNA, 21U-RNA, is found in *C. elegans*. These small RNAs are precisely 21 nucleotides and have a bias towards uridine at the 5′ end (but not in the remaining 20 nucleotides), and the genetic regions that encode them contain a characteristic sequence motif ~42 bp upstream of the first nucleotide of the small RNA[46]. It is possible that these RNAs are derived from thousands of separate, autonomously expressed, loci that are broadly scattered in two large regions of one chromosome. They are expressed solely in the germ line and interact with the PIWI-subfamily protein PRG-1 (refs 47, 48); therefore, 21U-RNAs are the *C. elegans* equivalent of piRNAs by definition. Like piRNAs, they depend on PRG-1 activity for their accumulation and are independent of DCR-1 (the *C. elegans* Dicer protein) for their production. *C. elegans* with mutations in *prg-1* have a smaller brood and a temperature-sensitive sterile phenotype, which is consistent with the idea that PIWI-subfamily proteins are involved in germline maintenance. Like the piRNAs found in mammalian germ cells in pachytene[33,34], 21U-RNAs have remarkable sequence diversity but lack obvious targets.

Small RNAs with a similar role to piRNAs have also been found in the ciliate *Tetrahymena thermophila*. These scan RNAs (scnRNAs) direct the elimination of transposon-like DNA sequences and associate with a PIWI-subfamily protein, TWI1 (ref. 8) but, in contrast to piRNAs and 21U-RNAs, are produced by a Dicer-dependent pathway[49].

These three examples (piRNAs, 21U-RNAs and scnRNAs) indicate that the core PIWI and piRNA machinery might have evolved to produce small RNAs and silence targets by different strategies.

RNA silencing pathways include mechanisms that downregulate endogenous genes and restrain the expression of selfish or exogenous genetic material, and these pathways often share common components such as Dicer. Therefore, there should be competition between different silencing pathways for particular components. Ways to overcome such competition should also exist; for example, by amplifying a weak silencing signal. In *C. elegans*, distinct Argonaute proteins operate at different stages of RNAi, directing gene silencing in a sequential manner[50] — the second stage of which involves RNaseIII-independent biogenesis of small RNAs. First, a primary Argonaute protein (such as RDE-1 for exogenous siRNAs (exo-siRNAs) and ERGO-1 for endo-siRNAs) is guided by 'primary' siRNAs (that is, a first round of siRNAs), which have been generated from long dsRNAs by DCR-1. Second, the silencing signal is amplified by the production of 'secondary' siRNAs by the action of RNA-dependent RNA polymerases (RdRPs) (Fig. 1). These secondary siRNAs then bind differentially to secondary Argonaute proteins (SAGOs, members of the WAGO subfamily), which mediate downstream silencing. In plants, RNAs with aberrant features, including lack of a poly(A) tail and lack of a 5′ cap, are copied into double-stranded forms by RdRPs and become substrates for Dicer, which converts them into siRNA duplexes[12]. By contrast, the *C. elegans* somatic RdRP mostly produces 21-nucleotide, single-stranded, 5′-triphosphorylated small RNAs directly from the target mRNA in a primer-independent manner without the need for Dicer-mediated cleavage of dsRNA[51–53]. Such recruitment of an RdRP directly to the target mRNA allows dsRNA synthesis without consuming the siRNAs generated in response to the original trigger, although it is unclear how the 3′ end of these secondary siRNAs is formed and what the molecular ruler is that determines their size.

Blurring of the boundaries between small RNA types

As described above, the three main classes of small RNA — siRNAs, miRNAs and piRNAs — are distinct in their biogenesis and cellular roles. However, recent findings blur these distinctions and show that there are even more-complex interactions between factors involved in small RNA biogenesis. Deep sequencing of small RNAs from somatic tissues and cultured somatic cells in *D. melanogaster* has uncovered another class of small RNA, consisting of 3′-methylated, 21-nucleotide RNAs derived from the *D. melanogaster* genome. These endogenous RNAs are derived from transposons and from several loci, including

loci that encode *cis*-natural antisense transcript pairs, and long stem–loop structures containing many mismatched pairs in their stems[54–57]. In *D. melanogaster*, distinct Dicer-containing complexes produce exo-siRNAs and miRNAs[58,59]. DCR-1 generates miRNAs, acting with its dsRNA-binding protein partner, Loquacious (LOQS)[60,61], and the miRNAs are loaded onto AGO1. By contrast, DCR-2, together with its dsRNA-binding protein partner, R2D2 (ref. 62), generates exo-siRNAs, which are loaded onto AGO2. Like exo-siRNAs, the recently discovered endogenous small RNAs are produced by the DCR-2-dependent pathway and are loaded onto AGO2, and they are therefore called endo-siRNAs. However, the generation of many endo-siRNAs requires LOQS[54,56], the dsRBD-containing partner of DCR-1 in the miRNA pathway[60,61], but not R2D2, the partner of DCR-2 (ref. 62). In *D. melanogaster* deficient in DCR-2 or AGO2, the expression of transposons increases, so endo-siRNAs might be the main mechanism for silencing 'selfish' genetic elements in somatic cells, which lack the piRNA pathway. Therefore, endo-siRNAs and piRNAs are fundamentally similar in that they defend organisms against nucleic-acid-based 'parasites'. This finding also shows that *D. melanogaster* has two RNAi pathways that repress transposon expression. Mouse oocytes have also been shown to contain endo-siRNAs. These RNAs are derived from various sources, including transposons[63,64]; however, some are processed from overlapping regions of functional genes and their cognate pseudogenes. This finding suggests that pseudogenes, which have been thought to be non-functional protein 'fossils', might regulate the expression of their founder genes.

Although siRNAs and miRNAs are categorized in terms of their origin rather than their size or function[7,12], the discovery of endo-siRNAs makes it difficult to distinguish between siRNAs and miRNAs. This blurring of the boundaries between the different types of small RNA has interesting evolutionary implications. The long stem–loop structures that are processed to form endo-siRNAs are reminiscent of the pre-miRNAs in plants. One hypothesis for the evolutionary origin of plant miRNAs is that new plant miRNA loci might evolve from the inverted duplication of founder loci, which when transcribed would result in hairpin RNAs[12]. These hairpin RNAs would have almost perfect self-complementarity and might be processed by Dicer-like enzymes other than DCL1, the main miRNA-processing enzyme in plants, because DCL1 has limited activity against such substrates. Subsequent acquisition of mutations as a result of genetic drift would produce a hairpin with imperfect complementarity, which could then be processed by DCL1. Thus, the stem–loop structures from which endo-siRNAs are derived could be evolutionary intermediates that are gradually transformed into miRNA precursors. It is possible that such an adaptive switch could also occur during the evolution of miRNA-encoding genes in *D. melanogaster*, in which DCR-1 would then generate miRNAs instead of endo-siRNAs being generated by DCR-2.

Loading and sorting of small RNAs by the RISC

In gene silencing pathways initiated by dsRNA precursors, Dicer-mediated cleavage yields small dsRNA intermediates (small RNA duplexes). These small RNA duplexes must be dissociated into 'competent' single strands in order to function as guides for RISCs. For each small RNA duplex, only one strand, the guide strand, is loaded onto a specific Argonaute protein and assembled into the active RISC; the other strand, the passenger strand, is destroyed. Many eukaryotes express more than one Argonaute protein, and these proteins bind to small RNAs in a sequence-independent manner. So how are small RNAs sorted and loaded onto a specific Argonaute protein?

Loading

A small RNA generated from dsRNA precursors is converted from a duplex into a single-stranded form as it is loaded into the RISC. The key steps in converting the RISC from its precursor form (the pre-RISC), which contains the small RNA duplex, to its mature form (the holo-RISC), which contains the guide strand, are small RNA strand unwinding and preferential strand selection. The prevalent view of RISC loading is that thermodynamic asymmetry along small RNA duplex determines which RNA strand is retained and which is discarded. More specifically, the strand that has its 5′ end at the thermodynamically less stable end of the small RNA duplex is preferentially loaded into the RISC as the guide strand, a phenomenon referred to as the asymmetry rule[65,66].

For siRNAs, the known interactions between Dicer and the Argonaute proteins[8] indicate that the production of the small RNA and the assembly of the RISC might be physically coupled. For example, in *D. melanogaster*, DCR-2 does not simply transfer siRNAs to a distinct RISC but, instead, forms part of the RISC together with the siRNAs, indicating that the role of DCR-2 extends beyond the initiation phase. The loading of siRNA duplexes onto AGO2 is facilitated by the RISC-loading complex, which contains DCR-2 and its dsRBD-containing partner, R2D2 (refs 62, 67). The particular strand of the siRNA duplex that is loaded onto AGO2 seems to be determined by the orientation of the DCR-2–R2D2 heterodimer on the siRNA duplex[68]. R2D2 is thought to sense the thermodynamic stability of the siRNA duplexes and bind to the more stable end of the siRNA, whereas DCR-2 is recruited to the less stable end. The heterodimer probably recruits AGO2 through an interaction between DCR-2 and AGO2. Previous models have proposed that the transition from a double-stranded silencing trigger to a single-stranded one is mediated by an unidentified ATP-dependent RNA helicase. However, the unwinding of the siRNA duplex and the loading of a single strand into the RISC are facilitated by the slicing of the unincorporated (passenger) strand by AGO2, a process that does not require ATP[69–71] (Fig. 1). Cleavage in the middle of the passenger strand, as though the passenger strand were an mRNA target, would be expected to reduce the annealing temperature and the free energy of duplex formation, which in turn facilitates the separation of the siRNA strands. These data support a model in which siRNAs are initially loaded as duplexes onto an AGO2-containing pre-RISC (Fig. 2).

By contrast, in humans, pre-miRNAs are known to bind to a preformed trimeric complex of AGO2, DICER1 and DICER1's dsRBD-containing partner, TRBP[72]. This complex can cleave target RNAs using pre-miRNA and can distinguish miRNA from miRNA*, in the absence of ATP hydrolysis[72,73], suggesting that DICER1-mediated cleavage and sensing of thermodynamic stability occur in series in the AGO2–DICER1–TRBP complex.

This process by which a pre-RISC is converted to a holo-RISC can also occur by a slicer-independent mechanism. Three of the four Argonaute proteins in humans (AGO1, AGO3 and AGO4) lack slicer activity but are nonetheless loaded with single-stranded guide siRNAs[9,11,28]. Similarly, single-stranded miRNAs are found associated with AGO2 in humans, despite the expectation that mismatches in the unwound pre-miRNA should block the passenger-strand cleavage activity of AGO2. Thus, a cleavage-independent (bypass) mechanism for RISC assembly must exist. RNA helicase A has been identified as a candidate for unwinding the duplex in this process[74].

Sorting

Once assembled, RISCs mediate a range of the effector steps in all RNA silencing mechanisms, from repressing translation to maintaining genome stability. The specialized functions of RISCs are likely to result from the particular proteins that associate with each Argonaute protein. In other words, the different RISC variants are distinguished by their constituent Argonaute protein. Thus, it is crucial that a specific set of small guide RNAs is directed to a specific Argonaute protein. Analyses of how different types of small RNA are channelled to different Argonaute proteins show that there are multiple mechanisms: the determinants for small RNA sorting vary from the structure of the small RNA duplex to the identity of the 5′ nucleotide and the presence and extent of modifications to this nucleotide.

In *D. melanogaster*, pre-miRNAs are processed by DCR-1, whereas exo-siRNA duplexes are produced by DCR-2 from long dsRNAs[58] (Fig. 2a). Small RNAs then seem to be loaded onto either AGO1 or AGO2, depending on the structure of a small intermediate RNA duplex[75]. If the duplex has a bulge in the middle (frequently observed in miRNA

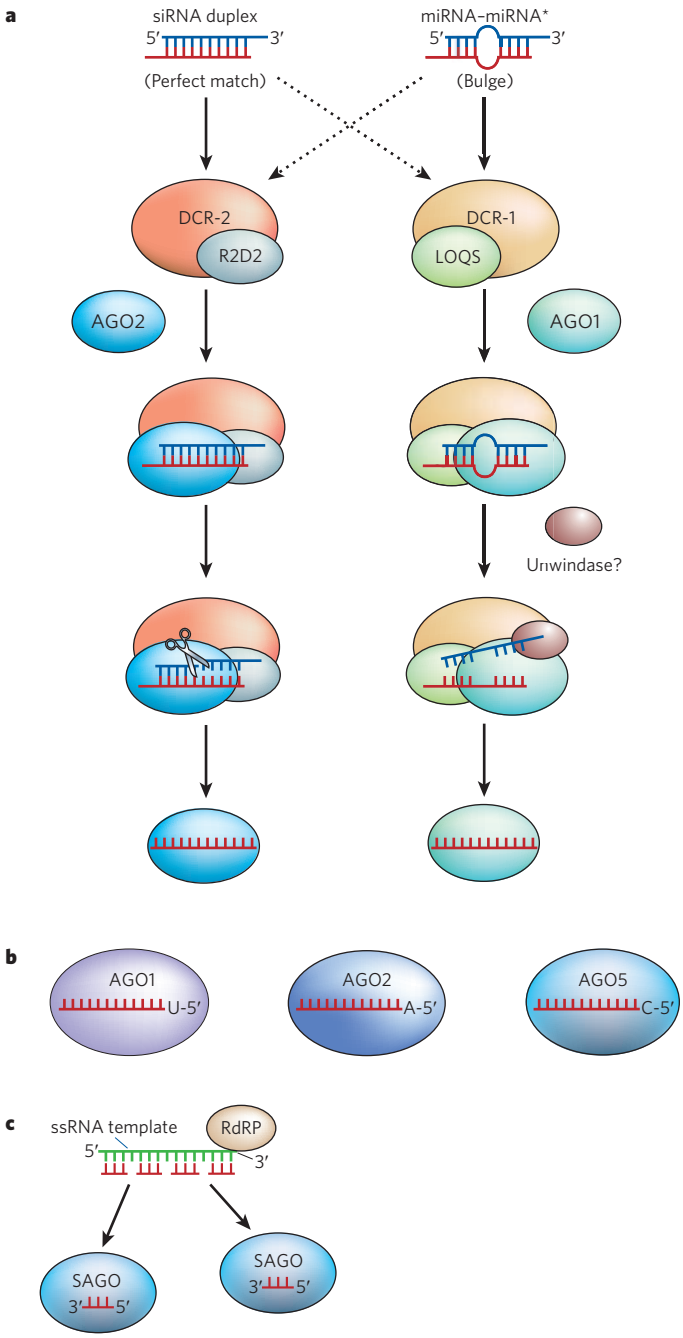


Figure 2 | Sorting of small RNAs onto distinct Argonaute proteins. Small RNAs are sorted onto specific Argonaute proteins, and this process occurs by several mechanisms. **a**, In *Drosophila melanogaster*, small RNAs originating from a duplex are loaded onto one of two Argonaute proteins (AGO1 or AGO2), on the basis of the structure of the small RNA duplex. If the duplex has a mismatch or a bulge in the centre (as miRNAs do), then the RNA is routed to AGO1. If the duplex is perfectly matched (as siRNAs are), then the small RNA is routed to AGO2. This selectivity occurs because the small RNAs are loaded onto Argonaute proteins from a Dicer-containing complex, and the two forms of Dicer, DCR-1 and DCR-2, associate with different RNA structures. DCR-2 pairs with R2D2, and this heterodimer binds to highly paired small RNA duplexes but recognizes small RNA duplexes with a central mismatch only poorly. AGO2 favours binding to DCR-2–R2D2 over binding to the other Dicer-containing complex, DCR-1–LOQS, which binds to small RNAs with bulges. Further processing into single-stranded small RNAs is described in Fig. 1. **b**, *Arabidopsis thaliana* miRNAs and *trans*-acting siRNAs (ta-siRNAs) have a 5′ uridine and preferentially associate with AGO1. By contrast, AGO2 and AGO5 show preferences for small RNAs containing 5′ adenosines and 5′ cytidines, respectively. However, it is unlikely that the 5′ nucleotide is the sole determinant of selective loading in *A. thaliana*. **c**, Secondary endo-siRNAs in *Caenorhabditis elegans*, as well in *Schizosaccharomyces pombe*, have a striking strand bias in which only the antisense siRNA is loaded onto Argonaute proteins. These siRNAs correspond to the RNA strand synthesized by RdRP. In *C. elegans*, RdRP produces small RNAs directly from the target mRNA in a primer-independent manner. Thus, these secondary small RNAs show negative polarity, and this mechanism reinforces the silencing carried out by the primary small RNAs.

precursors), the small RNA is routed to AGO1. If the duplex is perfectly matched, the small RNA is channelled to AGO2. This is because the DCR-2–R2D2 heterodimer, which recruits AGO2 to form the pre-RISC, binds well to highly paired small RNA duplexes but poorly to duplexes with central mismatches. Thus, the DCR-2–R2D2 heterodimer not only determines the polarity of siRNA loading on the basis of thermodynamic stability rules but also functions as a gatekeeper for AGO2-containing RISC assembly, promoting the incorporation of siRNAs over miRNAs. These observations suggest that each siRNA duplex dissociates from the active site of the Dicer protein after it is produced and is subsequently recaptured by the DCR-2–R2D2 heterodimer. However, although AGO1 favours binding to small RNA duplexes with central mismatches, a large proportion of miRNA–miRNA* duplexes with a base-paired central region still enter into AGO1-containing RISCs[55], suggesting that the AGO1-loading pathway is selective and not a default pathway for small RNAs rejected by the AGO2 pathway.

The identity of the nucleotide at the 5′ end and the extent to which this nucleotide is phosphorylated also influence which Argonaute protein the small RNA associates with. In contrast to what is observed in *D. melanogaster*, processing by Dicer may be uncoupled from association with Argonaute proteins in *A. thaliana* because, in this species, the miRNAs are all generated by one particular Dicer protein, DCL1, but are still sorted and loaded onto different Argonaute proteins. In *A. thaliana*, miRNAs and *trans*-acting siRNAs (ta-siRNAs), a class of small RNAs that regulate plant development[12], generally have a 5′ uridine and preferentially associate with AGO1 (ref. 76) (Fig. 2b). By contrast, AGO2 associates preferentially with small RNAs containing 5′ adenosines, and AGO5 prefers 5′ cytidines. Interestingly, if the opposite strand of a miRNA (that is, miRNA*) has a 5′ adenosine or a 5′ cytidine, it is bound to AGO2 or AGO5, respectively. These findings have led to the hypothesis that the binding affinity of Argonaute proteins for small RNAs is determined by the nucleotide at the 5′ end. Although these 5′-nucleotide preferences generally hold true for these Argonaute proteins in plants, exceptions have been observed: the *A. thaliana* miRNA known as miR-172 has a 5′ adenosine but preferentially associates with AGO1 (ref. 77); and AGO7 preferentially associates with

miR-390, which has a 5′ adenosine[77]. Therefore, the 5′ nucleotide does not seem to be the sole determinant of Argonaute association.

Another mechanism might operate for secondary siRNAs in *C. elegans*. These small RNAs are specifically loaded onto SAGOs[50]. Secondary siRNAs carry a 5′-triphosphate modification[51,52], the hallmark of RdRP products, which might function as a recognition element for SAGO binding while excluding binding by a primary Argonaute, such as RDE-1.

Endo-siRNAs in *C. elegans* (including the secondary siRNAs just mentioned) and *Schizosaccharomyces pombe* (fission yeast) have a striking strand bias in which only the antisense siRNA strand, corresponding to the RNA strand synthesized by RdRP, is loaded into Argonaute-containing complexes. Because *C. elegans* RdRPs produce small RNAs directly from the target mRNA, in a primer-independent manner (Fig. 2c), all secondary siRNAs have a negative polarity and function to reinforce the silencing of the target mRNA[50–52]. In *S. pombe*, the strand bias is probably the result of a different mechanism. The physical association of Dicer with an RdRP-containing complex known as RDRC and an Argonaute-containing complex known as the RNA-induced transcriptional silencing complex (RITS) (see page 413) may facilitate the loading of siRNAs onto Argonaute proteins in a directional manner as Dicer moves along and cleaves the dsRNA products of RdRP, giving rise to an antisense strand bias. This suggests that the polarity of Dicer processing defines the polarity of the siRNA strand loaded onto the Argonaute protein.

Argonaute proteins have diversified over evolutionary timescales, evolving a range of functions[8,12]. These findings about small RNA sorting imply that the diversification of the Argonaute proteins is a consequence of which small RNA they recruit. It is possible that the conformation of the Argonaute protein dictates which small RNAs it partners, but the structures of eukaryotic Argonaute proteins will need to be determined before this can be assessed.

Safeguards in silencing pathways

During RNA silencing, a single non-sequence-specific RNA-binding protein (Argonaute) is loaded with small guide RNAs with a variety of sequences, resulting in effector complexes (RISCs). Thus, this system requires gatekeepers to ensure that Argonaute can bind to small guide RNAs but not to degraded small RNAs, thereby avoiding 'off-target' silencing. Such gatekeeper systems seem to depend mainly on structural features specific for small guide RNAs.

As described earlier, Dicer helps to load siRNAs into the RISC, preventing siRNAs from diffusing freely in the cytoplasm after their production. This function of Dicer probably also aids in the discrimination of genuine siRNAs from various RNA-degradation products in the cell. Processing by RNaseIII enzymes (such as Dicer) characteristically yields small RNAs with 5′ monophosphates and 3′ two-nucleotide overhangs. The PAZ domain of Argonaute proteins might, as a first step, distinguish degraded RNAs (derived from unrelated pathways) from these small RNAs by binding to the characteristic 3′ overhangs of the small RNAs[8,12]. In addition, to become incorporated into the RISC and mediate cleavage of the target mRNA, the guide strand of an siRNA must have a phosphate group at the 5′ end[78]. In humans, the 5′ end of siRNAs is phosphorylated by the enzyme CLP1 (ref. 79), which also has roles in splicing transfer RNAs and forming the 3′ ends of mRNAs. Interestingly, both tRNA splicing and mRNA 3′-end formation occur in the nucleus[80,81], suggesting that siRNA duplexes with a 5′ hydroxyl group are transported to, or diffuse into, the nucleus and, after phosphorylation by CLP1, are exported to the cytoplasm and assembled into the RISC.

Amplification of the silencing signal needs to be balanced against the dangers of amplifying off-target silencing. For example, the slicer-mediated ping-pong mechanism for piRNA production does not lead to 'transitive' RNA silencing (in which RdRPs synthesize siRNAs complementary to sequences upstream or downstream of the initial trigger region in the target mRNA). Instead, it leads to conservative amplification of functional primary piRNA sequences (those inherited by germline transmission). However, it is conceivable that any off-target events mediated by RdRPs could lead to a chain reaction or transitive effect of silencing with deleterious consequences. Thus, there must be safeguards to prevent the pervasive use of RdRPs. A striking aspect of RdRP-based trigger amplification is that amplification occurs only when a target has been engaged, so amplification of the silencing signal is limited to cases in which there is a real target[51,52]. In *C. elegans*, the processing of the trigger dsRNA and the loading of primary siRNAs into the RDE-1-containing complex seem to be inherently inefficient, limiting the first round of target recognition by RDE-1-containing complexes and minimizing the risk of amplifying off-target silencing reactions[50]. In addition, each secondary siRNA seems to be generated by non-processive self-termination by RdRP, thereby restricting transitive effects[51–53]. Furthermore, secondary siRNAs associate with SAGOs, which lack catalytic residues for cleaving mRNAs, suggesting that these complexes cannot generate cleaved substrates for further amplification, which in turn would prevent them from inducing the exponential generation of secondary siRNAs[50] (but see also ref. 53 for a conflicting viewpoint). SAGOs are also present in limited supply and thus have a restricted capacity to support multiple simultaneous silencing reactions.

Another factor is that in *C. elegans* and *S. pombe* the RNAi machinery is negatively regulated by a conserved siRNA nuclease called enhanced RNAi (ERI-1 and Eri1, respectively)[82,83]. In *S. pombe*, transgene silencing is linked to a protein complex resembling the TRAMP complex of *Saccharomyces cerevisiae* (budding yeast), which carries out surveillance in the nucleus, targeting aberrant transcripts for degradation by the exosome[84]. Thus, RNAi in *S. pombe* is actively restricted from exerting its effects throughout the genome and seems to be subject to competition from RNA quality-control machinery.

Target-sensing modes and effector modes of the RISC

When the RISC is loaded with the guide strand of a small RNA, how does it find its target mRNA? Most of the binding energy that tethers a RISC to a target mRNA is from nucleotides in the seed region of the small RNA[85]. It seems that the accessibility of the target site can be sensed by the intrinsic, nonspecific affinity of RISC for ssRNA, which follows the initial specific association between the RISC and the target (through the 5′ seed region of the small RNA)[86]. But the accessibility of the target site correlates directly with the efficiency of cleavage, indicating that the RISC cannot unfold structured RNA.

Target mRNAs are present in the cell in complex with ribonucleoproteins (RNPs)[87], so target accessibility is also controlled by several RNA-binding proteins that either mask the target binding site or facilitate unfolding of the target. Therefore, the function of a RISC seems to be context-dependent, with its effector mode influenced not only by the structures of the small-RNA-binding sites on the target but also by the particular proteins associated with each Argonaute protein. For example, animal miRNAs silence gene expression by at least three independent mechanisms through binding sites that are mostly in the 3′ untranslated region of target mRNAs: by cleaving mRNAs, by repressing their translation and/or by promoting mRNA degradation[88,89]. However, the contribution of translational repression or mRNA degradation to gene silencing seems to differ for each miRNA–mRNA pair. Thus, the final outcome of miRNA regulation is probably affected by other proteins interacting with the targeted mRNA or RISC and counteracting the effects of the miRNA, resulting in differential regulation depending on the proteins present in each tissue[90].

Regulation of silencing pathways

So far, the pictures of RNA silencing pathways that we have built up (shown in Figs 1 and 2) are static. To gain further insight into silencing processes, it is important to incorporate information about how these pathways are regulated. It is already clear that competition between different silencing pathways (for example, competition between endo-siRNAs and miRNAs for LOQS in *D. melanogaster*) is a key step in how each stage of the RNAi mechanism is regulated. Many plant and animal viruses are known to encode suppressor proteins that block

host RNAi, and therefore silencing, at various stages[91]. Cellular proteins can also regulate RNAi. For example, processing to form the human miRNA let-7, which is a tumour suppressor and cell-cycle regulator, is post-transcriptionally inhibited in embryonic cells by the pluripotency factor LIN28, which seems to block the microprocessor-complex-mediated cleavage of pri-let-7 and the Dicer-mediated processing of pre-let-7 in series[92,93]. By contrast, in humans, signalling mediated by the transforming growth factor-β (TGF-β) and bone morphogenetic protein (BMP) family of growth factors rapidly increases the production of mature miR-21 (which is oncogenic), by promoting the processing of pri-miR-21 into pre-miR-21 by DROSHA[94]. More specifically, TGF-β- and BMP-specific signal transducers of the SMAD family are recruited to pri-miR-21 in complex with the RNA helicase p68, a component of the microprocessor complex, facilitating the accumulation of pre-miRNA. In addition, heterogeneous nuclear RNP A1 (hnRNP A1), a well-known regulator of precursor mRNA splicing, also assists DROSHA to crop and release pre-miR-18 efficiently, perhaps by refolding the hairpin or by creating a cleavage site for DROSHA through direct binding to the pri-miRNA[95]. This implies that some hairpins within pri-miRNAs might form and be processed only after the binding of a protein with RNA chaperone activity.

The activity of the RISC can also be regulated. In *A. thaliana*, the non-protein-coding gene *IPS1* (*INDUCED BY PHOSPHATE STARVATION 1*) contains a motif with sequence complementarity to the phosphate-starvation-induced miRNA miR-399, but the pairing is interrupted by a mismatched loop at the expected miRNA cleavage site[96]. *IPS1* mRNA is not cleaved but, instead, sequesters miR-399. Thus, *IPS1* overexpression results in increased accumulation of the target of miR-399, *PHO2* mRNA. The idea of target mimicry introduces unanticipated complexity into the network of RNA-regulatory interactions and raises the possibility that a large number of mRNA-like non-coding RNAs recently identified in humans[97] could be attenuators of the regulation of small-RNA–Argonaute complexes.

Perspective

Recent studies hint that human cells contain a large number of small RNAs similar to miRNAs or siRNAs, with the potential to regulate the expression of almost all human genes. The future challenges in this field are clear. Many questions remain to be answered. How many types of small RNA are there? How are these small RNAs generated? What are their biological functions? How are these pathways regulated? One potential problem is that because many types of small RNA are modified at their 5′ and 3′ ends[98], it is unclear whether the current sequencing technologies are sampling the entire range of small RNAs present in cells. But next-generation sequencing technologies[13] should soon help to uncover the full range of small RNA molecules.

One major challenge will be to identify how specific RNA-binding proteins affect the final outcome of gene regulation by small RNAs, given that RNAs in a cell are usually associated with multiple proteins that regulate many aspects of gene expression. For example, genome-wide *in vivo* approaches using a combination of immunoprecipitation and high-throughput sequencing will be required to establish protein–mRNA interactions or RNP complex occupancy at certain regions of mRNA, where expression is suppressed.

Finally, changes in the activity and specificity of silencing pathways could create quantitative and qualitative genetic variation in gene expression, thereby generating new gene-expression networks. Such changes might have contributed to many processes, including human evolution[16]. Given that all vertebrates have almost exactly the same number of protein-coding genes and therefore cannot readily be distinguished in this way, it might be prophetic that the first small guide RNA to be identified, the *C. elegans* miRNA lin-4, has been found to regulate a gene involved in the timing of development[99,100]. In humans, unlike other mammals, the brain tissue of newborns continues to grow at a similar rate to that of the fetus. This is a good example of a change in developmental timing, and there is much speculation about whether changes in this rate contributed to the evolution of humans as a new species. ■

1. Fire, A. *et al.* Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391,** 806–811 (1998).
2. Stefani, G. & Slack, F. J. Small non-coding RNAs in animal development. *Nature Rev. Mol. Cell Biol.* **9,** 219–230 (2008).
3. Ding, S. W. & Voinnet, O. Antiviral immunity directed by small RNAs. *Cell* **130,** 413–426 (2007).
4. Girard, A. & Hannon, G. J. Conserved themes in small-RNA-mediated transposon control. *Trends Cell Biol.* **18,** 136–148 (2008).
5. Hobert, O. Common logic of transcription factor and microRNA action. *Trends Biochem. Sci.* **29,** 462–468 (2004).
6. Meister, G. & Tuschl, T. Mechanisms of gene silencing by double-stranded RNA. *Nature* **431,** 343–349 (2004).
7. Tomari, Y. & Zamore, P. D. Machines for RNAi. *Genes Dev.* **19,** 517–529 (2005).
8. Hutvagner, G. & Simard, M. J. Argonaute proteins: key players in RNA silencing. *Nature Rev. Mol. Cell Biol.* **9,** 22–32 (2008).
9. Liu, J. *et al.* Argonaute2 is the catalytic engine of mammalian RNAi. *Science* **305,** 1437–1441 (2004).
10. Song, J. J., Smith, S. K., Hannon, G. J. & Joshua-Tor, L. Crystal structure of Argonaute and its implications for RISC slicer activity. *Science* **305,** 1434–1437 (2004).
11. Meister, G. *et al.* Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. *Mol. Cell* **15,** 185–197 (2004).
12. Chapman, E. J. & Carrington, C. Specialization and evolution of endogenous small RNA pathways. *Nature Rev. Genet.* **8,** 884–896 (2007).
13. Mardis, E. R. The impact of next-generation sequencing technology on genetics. *Trends Genet.* **24,** 133–141 (2008).
14. Zhang, H., Kolb, F. A., Jaskiewicz, L., Westhof, E. & Filipowicz, W. Single processing center models for human Dicer and bacterial RNase III. *Cell* **118,** 57–68 (2004).
15. MacRae, I. J., Zhou, K. & Doudna, J. A. Structural determinants of RNA recognition and cleavage by Dicer. *Nature Struct. Mol. Biol.* **14,** 934–940 (2007).
16. Heimberg, A. M. *et al.* MicroRNAs and the advent of vertebrate morphological complexity. *Proc. Natl Acad. Sci. USA* **105,** 2946–2950 (2008).
17. Kim, V. N. MicroRNA biogenesis: coordinated cropping and dicing. *Nature Rev. Mol. Cell Biol.* **6,** 376–385 (2005).
18. Han, J. *et al.* Molecular basis for the recognition of primary microRNAs by the Drosha–DGCR8 complex. *Cell* **125,** 887–901 (2006).
19. Zeng, Y., Yi, R. & Cullen, B. Recognition and cleavage of primary microRNA precursors by the nuclear processing enzyme Drosha. *EMBO J.* **24,** 138–148 (2005).
20. Denli, A. M., Tops, B. B., Plasterk, R. H., Ketting, R. F. & Hannon, G. J. Processing of primary microRNAs by the Microprocessor complex. *Nature* **432,** 231–235 (2004).
21. Gregory, R. I. *et al.* The Microprocessor complex mediates the genesis of microRNAs. *Nature* **432,** 235–240 (2004).
22. Ruby, J. G., Jan, C. H. & Bartel, D. P. Intronic microRNA precursors that bypass Drosha processing. *Nature* **448,** 83–86 (2007).
23. Okamura, K., Hagen, J. W., Duan, H., Tyler, D. M. & Lai, E. C. The mirtron pathway generates microRNA-class regulatory RNAs in *Drosophila*. *Cell* **130,** 89–100 (2007).
24. Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120,** 15–20 (2005).
25. Brennecke, J., Stark, A., Russell, R. B. & Cohen, S. M. Principles of microRNA–target recognition. *PLoS Biol.* **3,** e85 (2005).
26. Landgraf, P. *et al.* A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell* **129,** 1401–1414 (2007).
27. Morin, R. D. *et al.* Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res.* **18,** 610–621 (2008).
28. Azuma-Mukai, A. *et al.* Characterization of endogenous human Argonautes and their miRNA partners in RNA silencing. *Proc. Natl Acad. Sci. USA* **105,** 7964–7969 (2008).
29. Faehnle, C. R. & Joshua-Tor, L. Argonautes confront new small RNAs. *Curr. Opin. Chem. Biol.* **11,** 569–577 (2007).
30. Wang, Y., Sheng, G., Juranek, S., Tuschl, T. & Patel, D. J. Structure of the guide-strand-containing argonaute silencing complex. *Nature* **456,** 209–213 (2008).
31. Vagin, V. V. *et al.* A distinct small RNA pathway silences selfish genetic elements in the germline. *Science* **313,** 320–324 (2006).
This paper was the first to describe a Dicer-independent pathway for the biogenesis of the guide strand of small RNAs.
32. Saito, K. *et al.* Specific association of Piwi with rasiRNAs derived from retrotransposon and heterochromatic regions in the *Drosophila* genome. *Genes Dev.* **20,** 2214–2222 (2006).
33. Girard, A., Sachidanandam, R., Hannon, G. J. & Carmell, M. A. A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature* **442,** 199–202 (2006).
34. Aravin, A. *et al.* A novel class of small RNAs bind to MILI protein in mouse testes. *Nature* **442,** 203–207 (2006).
35. Lau, N. C. *et al.* Characterization of the piRNA complex from rat testes. *Science* **313,** 363–367 (2006).
36. Watanabe, T. *et al.* Identification and characterization of two novel classes of small RNAs in the mouse germline: retrotransposon-derived siRNAs in oocytes and germline small RNAs in testes. *Genes Dev.* **20,** 1732–1743 (2006).
37. Yin, H. & Lin, H. An epigenetic activation role of Piwi and a Piwi-associated piRNA in *Drosophila melanogaster*. *Nature* **450,** 304–308 (2007).
38. Saito, K. *et al.* Pimet, the *Drosophila* homolog of HEN1, mediates 2′-O-methylation of Piwi-interacting RNAs at their 3′ ends. *Genes Dev.* **21,** 1603–1608 (2007).
39. Horwich, M. D. *et al.* The *Drosophila* RNA methyltransferase, DmHen1, modifies germline piRNAs and single-stranded siRNAs in RISC. *Curr. Biol.* **17,** 1265–1272 (2007).
40. Kirino, Y. & Mourelatos, Z. Mouse Piwi-interacting RNAs are 2′-O-methylated at their 3′ termini. *Nature Struct. Mol. Biol.* **14,** 347–348 (2007).
41. Ohara, T. *et al.* The 3′ termini of mouse Piwi-interacting RNAs are 2′-O-methylated. *Nature Struct. Mol. Biol.* **14,** 349–350 (2007).
42. Houwing, S. *et al.* A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in zebrafish. *Cell* **129,** 69–82 (2007).
43. Brennecke, J. *et al.* Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell* **128,** 1089–1103 (2007).

44. Gunawardane, L. S. *et al.* A slicer-mediated mechanism for repeat-associated siRNA 5′ end formation in *Drosophila*. *Science* **315,** 1587–1590 (2007).
References 43 and 44 were the first to describe slicer-mediated small RNA production.
45. Aravin, A. A., Sachidanandam, R., Girard, A., Fejes-Toth, K. & Hannon, G. J. Developmentally regulated piRNA clusters implicate MILI in transposon control. *Science* **316,** 744–747 (2007).
46. Ruby, J. G. *et al.* Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell* **127,** 1193–1207 (2006).
47. Batista, P. J. *et al.* PRG-1 and 21U-RNAs interact to form the piRNA complex required for fertility in *C. elegans*. *Mol. Cell* **31,** 67–78 (2008).
48. Das, P. P. *et al.* Piwi and piRNAs act upstream of an endogenous siRNA pathway to suppress Tc3 transposon mobility in the *Caenorhabditis elegans* germline. *Mol. Cell* **31,** 79–90 (2008).
49. Mochizuki, K. & Gorovsky, M. A. A Dicer-like protein in *Tetrahymena* has distinct functions in genome rearrangement, chromosome segregation, and meiotic prophase. *Genes Dev.* **19,** 77–89 (2005).
50. Yigit, E. *et al.* Analysis of the *C. elegans* Argonaute family reveals that distinct Argonautes act sequentially during RNAi. *Cell* **127,** 747–757 (2006).
This paper shows that RNAi occurs by a two-step pathway in *C. elegans*.
51. Pak, J. & Fire, A. Distinct populations of primary and secondary effectors during RNAi in *C. elegans*. *Science* **315,** 241–244 (2007).
52. Sijen, T., Steiner, F. A., Thijssen, K. L. & Plasterk, R. H. Secondary siRNAs result from unprimed RNA synthesis and form a distinct class. *Science* **315,** 244–247 (2007).
53. Aoki, K., Moriguchi, H., Yoshioka, T., Okawa, K. & Tabara, H. *In vitro* analyses of the production and activity of secondary small interfering RNAs in *C. elegans*. *EMBO J.* **26,** 5007–5019 (2007).
54. Czech, B. *et al.* An endogenous small interfering RNA pathway in *Drosophila*. *Nature* **453,** 798–802 (2008).
55. Kawamura, Y. *et al. Drosophila* endogenous small RNAs bind to Argonaute 2 in somatic cells. *Nature* **453,** 793–797 (2008).
56. Okamura, K. *et al.* The *Drosophila* hairpin RNA pathway generates endogenous short interfering RNAs. *Nature* **453,** 803–806 (2008).
57. Ghildiyal, M. *et al.* Endogenous siRNAs derived from transposons and mRNAs in *Drosophila* somatic cells. *Science* **320,** 1077–1081 (2008).
58. Lee, Y. S. *et al.* Distinct roles for *Drosophila* Dicer-1 and Dicer-2 in the siRNA/miRNA silencing pathways. Cell **117,** 69–81 (2004).
59. Okamura, K., Ishizuka, A., Siomi, H. & Siomi, M. C. Distinct roles for Argonaute proteins in small RNA-directed RNA cleavage pathways. *Genes Dev.* **18,** 1655–1666 (2004).
60. Saito, K., Ishizuka, A., Siomi, H. & Siomi, M. C. Processing of pre-microRNAs by the Dicer-1-Loquacious complex in *Drosophila* cells. *PLoS Biol.* **3,** e235 (2005).
61. Förstemann, K. *et al.* Normal microRNA maturation and germ-line stem cell maintenance requires Loquacious, a double-stranded RNA-binding domain protein. *PLoS Biol.* **3,** e236 (2005).
62. Liu, Q. *et al.* R2D2, a bridge between the initiation and effector steps of the *Drosophila* RNAi pathway. *Science* **301,** 1921–1925 (2003).
63. Tam, O. H. *et al.* Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* **453,** 534–538 (2008).
64. Watanabe, T. *et al.* Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* **453,** 539–543 (2008).
65. Schwarz, D. S. *et al.* Asymmetry in the assembly of the RNAi enzyme complex. *Cell* **115,** 199–208 (2003).
66. Khvorova, A., Reynolds, A. & Jayasena, S. D. Functional siRNAs and miRNAs exhibit strand bias. *Cell* **115,** 209–216 (2003).
67. Liu, X., Jiang, F., Kalidas, S., Smith, D. & Liu, Q. Dicer-2 and R2D2 coordinately bind siRNA to promote assembly of the siRISC complexes. *RNA* **12,** 1514–1520 (2006).
68. Tomari, Y., Matranga, C., Haley, B., Martinez, N. & Zamore, P. D. A protein sensor for siRNA asymmetry. *Science* **306,** 1377–1380 (2004).
69. Matranga, C., Tomari, Y., Shin, C., Bartel, D. P. & Zamore, P. D. Passenger-strand cleavage facilitates assembly of siRNA into Ago2-containing RNAi enzyme complexes. *Cell* **123,** 607–620 (2005).
70. Rand, T. A., Petersen, S., Du, F. & Wang, X. Argonaute2 cleaves the anti-guide strand of siRNA during RISC activation. *Cell* **123,** 621–629 (2005).
71. Miyoshi, K., Tsukumo, H., Nagami, T., Siomi, H. & Siomi, M. C. Slicer function of *Drosophila* Argonautes and its involvement in RISC formation. *Genes Dev.* **19,** 2837–2848 (2005).
72. Gregory, R. I., Chendrimada, T. P., Cooch, N. & Shiekhattar, R. Human RISC couples microRNA biogenesis and posttranscriptional gene silencing. *Cell* **123,** 631–640 (2005).
73. Maniataki, E. & Mourelatos, Z. A human, ATP-independent, RISC assembly machine fueled by pre-miRNA. *Genes Dev.* **19,** 2979–2990 (2005).
74. Robb, G. B. & Rana, T. M. RNA helicase A interacts with RISC in human cells and functions in RISC loading. *Mol. Cell* **26,** 523–537 (2007).
75. Tomari, Y., Du, T. & Zamore, P. D. Sorting of *Drosophila* small silencing RNAs. *Cell* **130,** 299–308 (2007).
76. Mi, S. *et al.* Sorting of small RNAs into *Arabidopsis* Argonaute complexes is directed by the 5′ terminal nucleotide. *Cell* **133,** 116–127 (2008).
77. Montgomery, T. A. *et al.* Specificity of ARGONAUTE7–miR390 interaction and dual functionality in *TAS3 trans*-acting siRNA formation. *Cell* **133,** 128–141 (2008).
References 76 and 77 show that the sorting of plant miRNAs onto Argonaute proteins depends mainly on the nucleotide at the 5′ end.
78. Pham, J. W. & Sontheimer, E. J. Molecular requirements for RNA-induced silencing complex assembly in the *Drosophila* RNA interference pathway. *J. Biol. Chem.* **280,** 39278–39283 (2005).
79. Weitzer, S. & Martinez, J. The human RNA kinase hClp1 is active on 3′ transfer RNA exons and short interfering RNAs. *Nature* **447,** 222–226 (2007).
80. Paushkin, S. V., Patel, M., Furia, B. S., Peltz, S. W. & Trotta, C. R. Identification of a human endonuclease complex reveals a link between tRNA splicing and pre-mRNA 3′ end formation. *Cell* **117,** 311–321 (2004).
81. Danckwardt, S., Hentze, M. W. & Kulozik, A. E. 3′ end mRNA processing: molecular mechanisms and implications for health and disease. *EMBO J.* **27,** 482–498 (2008).
82. Kennedy, S., Wang, D. & Ruvkun, G. A conserved siRNA-degrading RNase negatively regulates RNA interference in *C. elegans*. *Nature* **427,** 645–649 (2004).
83. Iida, T., Kawaguchi, R. & Nakayama, J. Conserved ribonuclease, Eri1, negatively regulates heterochromatin assembly in fission yeast. *Curr. Biol.* **16,** 1459–1464 (2006).
84. Bühler, M., Haas, W., Gygi, S. P. & Moazed, D. RNAi-dependent and -independent RNA turnover mechanisms contribute to heterochromatic gene silencing. *Cell* **129,** 707–721 (2007).
85. Haley, B. & Zamore, P. D. Kinetic analysis of the RNAi enzyme complex. *Nature Struct. Mol. Biol.* **11,** 599–606 (2004).
86. Ameres, S. L., Martinez, J. & Schroeder, R. Molecular basis for target RNA recognition and cleavage by human RISC. *Cell* **130,** 101–112 (2007).
87. Dreyfuss, G., Kim, V. N. & Kataoka, N. Messenger-RNA-binding proteins and the messages they carry. *Nature Rev. Mol. Cell Biol.* **3,** 195–205 (2002).
88. Filipowicz, W., Bhattacharyya, S. N. & Sonenberg, N. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nature Rev. Genet.* **9,** 102–114 (2008).
89. Eulalio, A., Behm-Ansmant, I. & Izaurralde, E. P bodies: at the crossroads of post-transcriptional pathways. *Nature Rev. Mol. Cell Biol.* **8,** 9–22 (2007).
90. Kedde, M. *et al.* RNA-binding protein Dnd1 inhibits microRNA access to target mRNA. *Cell* **131,** 1273–1286 (2007).
This paper shows that the final outcome of miRNA regulation is affected by the interaction of proteins other than Argonaute with the target mRNA.
91. Mlotshwa, S., Pruss, G. J. & Vance, V. Small RNAs in viral infection and host defense. *Trends Plant Sci.* **13,** 375–382 (2008).
92. Viswanathan, S. R., Daley, G. Q. & Gregory, R. I. Selective blockade of microRNA processing by Lin28. *Science* **320,** 97–100 (2008).
93. Rybak, A. *et al.* A feedback loop comprising *lin-28* and *let-7* controls pre-*let-7* maturation during neural stem-cell commitment. *Nature Cell Biol.* **10,** 987–993 (2008).
94. Davis, B. N., Hilyard, A. C., Lagna, G. & Hata, A. SMAD proteins control DROSHA-mediated microRNA maturation. *Nature* **454,** 56–61 (2008).
95. Guil, S. & Cáceres, J. F. The multifunctional RNA-binding protein hnRNP A1 is required for processing of miR-18a. *Nature Struct. Mol. Biol.* **14,** 591–596 (2007).
96. Franco-Zorrilla, J. M. *et al.* Target mimicry provides a new mechanism for regulation of microRNA activity. *Nature Genet.* **39,** 1033–1037 (2007).
This paper describes how the activity of miRNAs can be regulated by transcripts that mimic the target transcript.
97. ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447,** 799–816 (2007).
98. Farazi, T. A., Juranek, S. A. & Tuschl, T. The growing catalog of small RNAs and their association with distinct Argonaute/Piwi family members. *Development* **135,** 1201–1214 (2008).
99. Lee, R. C., Feinbaum, R. L. & Ambros, V. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* **75,** 843–854 (1993).
100. Wightman, B., Ha, I. & Ruvkun, G. Post-transcriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *C. elegans*. *Cell* **75,** 855–862 (1993).

Acknowledgements We apologize to colleagues whose relevant primary publications were not cited because of space constraints. We thank Y. Tomari, K. Aoki, Y. Watanabe and all the members of the Siomi laboratory for their comments and critical reading of the manuscript. Work in our laboratory is supported by grants from the Ministry of Education, Culture, Sports, Science and Technology (MEXT, Japan) and the New Energy and Industrial Technology Development Organization (Japan). M.C.S. is associate professor of the Global Centre of Excellence for Human Metabolomics Systems Biology (MEXT).

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence should be addressed to the authors (awa403@sc.itc.keio.ac.jp; siomim@sc.itc.keio.ac.jp).

A three-dimensional view of the molecular machinery of RNA interference

Martin Jinek[1] & Jennifer A. Doudna[1–4]†

In eukaryotes, small non-coding RNAs regulate gene expression, helping to control cellular metabolism, growth and differentiation, to maintain genome integrity, and to combat viruses and mobile genetic elements. These pathways involve two specialized ribonucleases that control the production and function of small regulatory RNAs. The enzyme Dicer cleaves double-stranded RNA precursors, generating short interfering RNAs and microRNAs in the cytoplasm. These small RNAs are transferred to Argonaute proteins, which guide the sequence-specific silencing of messenger RNAs that contain complementary sequences by either enzymatically cleaving the mRNA or repressing its translation. The molecular structures of Dicer and the Argonaute proteins, free and bound to small RNAs, have offered exciting insights into the molecular mechanisms that are central to RNA silencing pathways.

The discovery that RNA interference (RNAi) and related small-RNA-mediated pathways have central roles in the silencing of gene expression in eukaryotic cells has profoundly altered the understanding of gene regulation. At least 30% of human genes are thought to be regulated by microRNAs, one of the classes of small RNA[1]. In a wide range of organisms, including petunias, nematodes, fruitflies, zebrafish and mice, mutations in the genetic regions encoding the protein and/or RNA components of RNAi result in severe and sometimes lethal defects in cell growth and development[2]. These findings have raised many questions about how and why this widespread RNA-mediated regulation of genes evolved and how these mechanisms enable gene expression to be precisely tuned in response to internal and external stimuli.

RNAi and related gene-silencing pathways are initiated by the production of small RNAs (~20–30 nucleotides) with sequences that are complementary to portions of the transcripts that they regulate. There are three main classes of small regulatory RNA: short interfering RNAs (siRNAs), microRNAs (miRNAs) and PIWI-interacting RNAs (piRNAs) (see page 396 for a review of the currently recognized classes of small regulatory RNA). The biogenesis and mechanism of action of the main types of small RNA are described in Box 1. In brief, siRNAs and miRNAs are generated from double-stranded RNA (dsRNA) precursors that are produced in, or introduced into, cells, and their generation depends on the ribonuclease (RNase) Dicer[3]. These small RNAs subsequently associate with members of the Argonaute family of proteins, which function as the core components of a diverse set of protein–RNA complexes called RNA-induced silencing complexes (RISCs)[4]. RISCs use the small RNAs as guides for the sequence-specific silencing of messenger RNAs that contain complementary sequence through inducing the degradation of the mRNAs or repressing their translation. In addition, in certain organisms, a specialized nuclear Argonaute-containing complex, known as the RNA-induced transcriptional silencing complex (RITS), mediates transcriptional gene silencing by inducing heterochromatin formation[5] (see page 413). The biogenesis of the piRNA class of small RNAs also involves proteins belonging to the Argonaute family but differs markedly from that of siRNAs and miRNAs[6] (Box 1).

Biochemical and structural biology studies have provided fundamental insights into the molecular details of RNAi and its possible evolutionary underpinnings. Structural aspects of how viruses inhibit host-cell RNAi pathways are described in refs 7 and 8. In this Review, we focus on two specialized RNases: Dicer, which functions as a molecular ruler in siRNA and miRNA biogenesis; and Argonaute, a versatile RNA-guided molecular machine that cleaves, or otherwise represses, target RNAs. We highlight the ways in which recently obtained molecular structures of these two proteins underlie the current mechanistic understanding of RNA silencing.

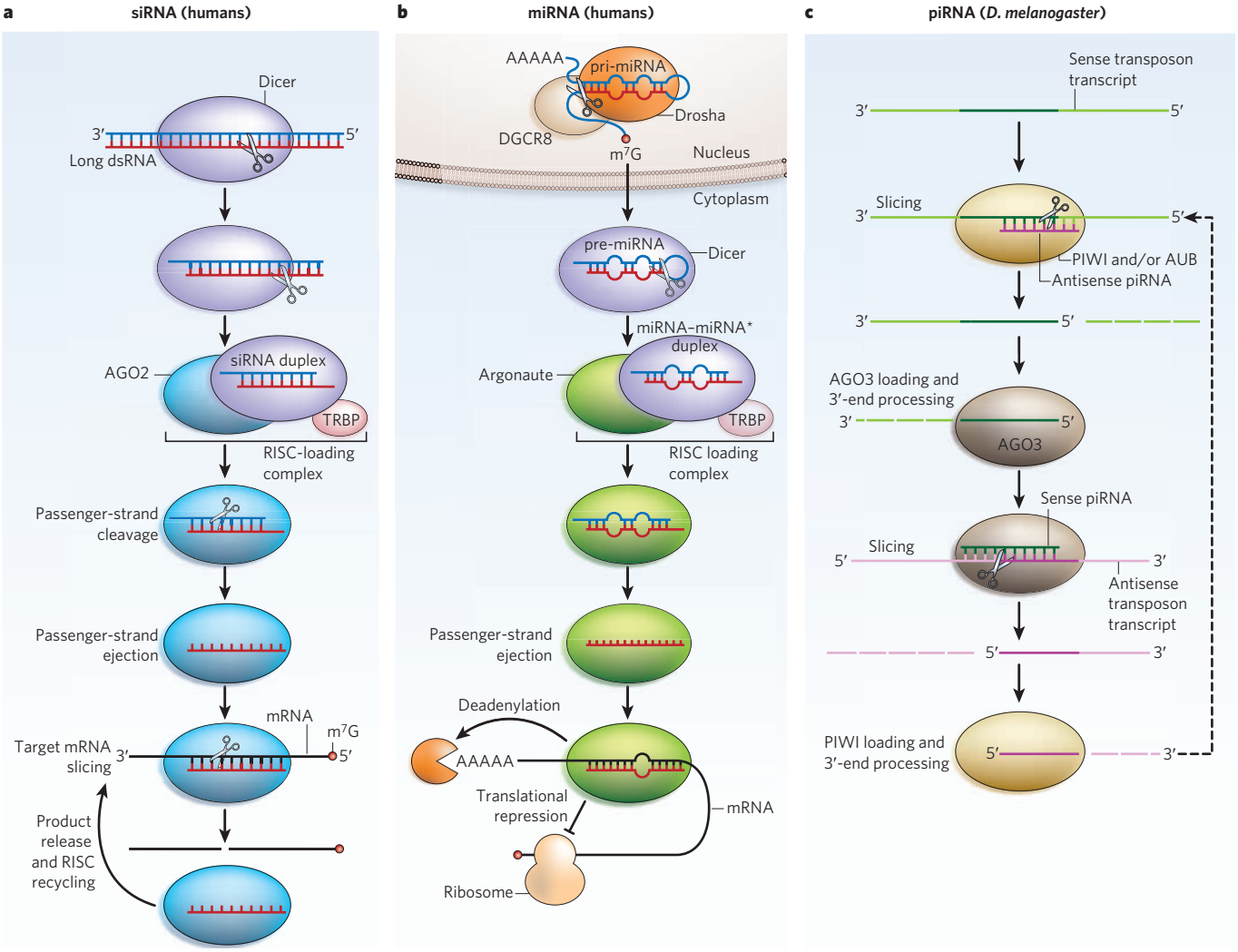
Structural aspects of small RNA biogenesis

The production of siRNAs and miRNAs relies on the endonucleolytic processing of dsRNA precursors. In a cell, long dsRNAs can arise from the replication of RNA viruses, from the transcription of convergent cellular genes or mobile genetic elements, and from self-annealing cellular transcripts. Dicer, an endonuclease belonging to the RNaseIII family, processes these long dsRNA molecules to yield siRNA duplexes of ~21–25 nucleotides[3]. By contrast, miRNAs are generated from endogenous transcripts (known as primary miRNAs, pri-miRNAs) that form stem–loop structures (Box 1). The hairpin region is excised from the pri-miRNA in the nucleus by the endonuclease Drosha[9], another RNaseIII-family enzyme. After its export to the cytoplasm, the hairpin (known as a precursor miRNA, pre-miRNA) undergoes another endonucleolytic cleavage, which is catalysed by Dicer, generating a miRNA–miRNA* duplex (where miRNA is the antisense, or guide, strand and miRNA* is the sense, or passenger, strand) of ~21–25 nucleotides.

Two features of siRNAs and miRNAs ensure that they are efficiently incorporated into the RISC: the length of the duplex; and characteristic 5′ and 3′ ends, carrying a monophosphate group and a dinucleotide overhang, respectively[10–12]. Recent structural studies of a prokaryotic RNaseIII and a eukaryotic Dicer have provided insights into how these crucial features of siRNAs and miRNAs are generated during their biogenesis by Drosha and/or Dicer.

[1]Department of Molecular and Cell Biology, [2]Howard Hughes Medical Institute, [3]Department of Chemistry, University of California, Berkeley, California 94720, USA. [4]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. †Present address: Genentech, 1 DNA Way, South San Francisco, California 94080, USA.

Box 1 | Biogenesis and mechanism of action of the main classes of small regulatory RNA



Small regulatory RNAs are non-coding RNA molecules that silence target RNAs in a sequence-specific manner. The main classes of small RNA are short interfering RNAs (siRNAs), microRNAs (miRNAs) and PIWI-interacting RNAs (piRNAs).

The first class, siRNAs, mediate RNAi by downregulating target RNAs through slicing; that is, by endonucleolytic cleavage (see figure, panel **a**; note that RNAs are not drawn to the same scale in each panel, and m^7G denotes 7-methylguanosine). These small RNAs are derived from long double-stranded RNA (dsRNA) molecules that result from RNA virus replication, convergent transcription of cellular genes or mobile genetic elements, self-annealing transcripts or experimental transfection. The endonuclease Dicer functions as a molecular ruler to cleave the dsRNA at ~21–25-nucleotide intervals. After Dicer-mediated cleavage, one strand of the siRNA duplex (the guide strand) is loaded onto an Argonaute protein at the core of an RNA-induced silencing complex (RISC). (For simplicity, in the figure, the RISC is represented just by an Argonaute protein.) Argonaute loading takes place in the RISC-loading complex, a ternary complex that consists of an Argonaute protein, Dicer and a dsRNA-binding protein (known as TRBP in humans). During loading, the non-guide (passenger) strand is cleaved by an Argonaute protein and ejected. The Argonaute protein then uses the guide siRNA to associate with target RNAs that contain perfectly complementary sequence and then catalyses the slicing of these targets. After slicing, the cleaved target RNA is released, and the RISC is recycled for another round of slicing.

The second class, miRNAs, are encoded in the genome. Whereas plant miRNAs direct the slicing of target messenger RNAs, much like siRNAs, animal miRNAs silence target mRNAs without slicing (panel **b**). These small RNAs are transcribed from endogenous miRNA genes as primary transcripts (pri-miRNAs), containing ~65–70-nucleotide stem–loop structures. The hairpin structure is excised in the nucleus by the Drosha–DGCR8 complex to yield a precursor miRNA (pre-miRNA). In the cytoplasm, Dicer cleaves the pre-miRNA, producing a miRNA–miRNA* duplex (where miRNA denotes the guide strand and miRNA* the passenger strand). The guide strand is loaded onto an Argonaute protein. Typically, animal miRNAs are only partly complementary to sequences in the 3′ untranslated regions of their target mRNAs, and this lack of complementarity prevents the target from being sliced by the Argonaute protein. In addition, some Argonaute proteins involved in the miRNA pathway lack the catalytic residues needed for slicer activity. The mechanism of miRNA-mediated silencing is unresolved but is thought to occur by repression of target mRNA translation and removal of mRNA poly(A) tails (that is, deadenylation), which leads to mRNA degradation.

The third class, piRNAs, are ~24–31 nucleotides, and they silence transposons (mobile genetic elements) in animal germ cells. The biogenesis and mechanism of action of piRNAs is poorly understood. A model for these processes in *Drosophila melanogaster* is shown in panel **c**. The precursors of piRNAs are single-stranded RNAs, because piRNA biogenesis does not require Dicer. The small RNAs induce reciprocal slicer-dependent cleavages of sense and antisense transposon transcripts. This process is mediated by the PIWI clade of the Argonaute family, which includes the proteins PIWI, Aubergine (AUB) and Argonaute 3 (AGO3) in *D. melanogaster*. PIWI- or AUB-mediated slicing of sense transcripts generates sense piRNAs, which associate with AGO3 and direct the slicing of antisense transposon transcripts. The slicing products give rise to antisense piRNAs, which in turn bind to PIWI and AUB and guide the slicing of sense transposon transcripts to generate sense piRNAs.

The RNaseIII family

Cleavage of dsRNA by enzymes of the RNaseIII family, including Drosha and Dicer, yields products with characteristic termini, with a monophosphate group at the 5′ ends, and a two-nucleotide overhang at the 3′ ends[13]. The simplest RNaseIII enzymes are found in prokaryotes and certain fungi. These enzymes consist of an RNaseIII domain, which has the catalytic activity, and (generally) a dsRNA-binding domain (dsRBD) (Fig. 1a), and they function as homodimers[14]. The two RNaseIII domains of the dimer associate to form a single processing centre, with each catalytic domain responsible for the hydrolysis of one strand in the duplex[15]. By contrast, both Drosha and Dicer are monomeric and contain two tandemly arranged RNaseIII domains and a single dsRBD.

From the crystal structure of RNaseIII from the prokaryote *Aquifex aeolicus* in complex with a cleaved dsRNA product[16] (Fig. 1b), it can be seen that homodimerization of the catalytic domains creates a shallow surface cleft separating the two active sites. A cluster of conserved acidic amino-acid residues in each catalytic centre coordinates a single magnesium ion. On substrate binding, the enzyme rearranges such that the dsRBDs clamp the dsRNA substrate over the surface of the catalytic domain dimer.

Dicer

The endonuclease Dicer processes dsRNA substrates (long dsRNAs and pre-miRNAs) into short dsRNA fragments (siRNAs and miRNAs) of defined length, typically ~21–25 nucleotides[3]. In addition to two copies of the conserved RNaseIII domain and a dsRBD in the carboxyl terminus, Dicer enzymes usually have an amino-terminal DEXD/H-box domain, followed by a small domain of unknown function (the DUF283 domain) and a PAZ domain (Fig. 1a). The PAZ domain, which is also found in Argonaute proteins, binds specifically to the 3′ end of single-stranded RNA (ssRNA)[17–19].

The crystal structure of Dicer from the unicellular eukaryote *Giardia intestinalis* revealed that the ability of Dicer enzymes to produce dsRNA fragments of specific length originates from a unique spatial arrangement of the PAZ domain and the RNaseIII domains[20]. *G. intestinalis* Dicer is a naturally 'trimmed-down' version of the enzyme, consisting only of the PAZ domain and the tandem RNaseIII domains (RNaseIIIa and RNaseIIIb) (Fig. 1a). Its structure resembles an axe, with the two RNaseIII catalytic domains forming the blade and the PAZ domain making up the base of the handle (Fig. 1c). The RNaseIIIa domain and the PAZ domain are connected by a long helix running the length of the handle. This connector helix is buttressed by a platform domain formed by the N-terminal segment of the protein. The RNaseIII domains form an intramolecular dimer that closely resembles the homodimeric structure of prokaryotic RNaseIII[16] (Fig. 1b). In Dicer, four conserved acidic amino-acid residues in the active site of each RNaseIII domain coordinate two metal cations, suggesting that Dicer uses a two-metal-ion mechanism to catalyse RNA cleavage. The 17.5 Å distance between the two metal-ion pairs in the two active sites matches the width of the major groove in a dsRNA duplex. Modelling the binding of a dsRNA substrate to the enzyme reveals that the duplex runs along a flat surface formed by the platform domain and makes electrostatic interactions with a number of positively charged residues. Mutation of the sequence encoding these residues impairs the catalytic activity of Dicer, underscoring the importance of these positively charged residues for substrate binding[21].

The PAZ domain of Dicer has the same fold and 3′-overhang-binding residues as the PAZ domain of Argonaute proteins (discussed in the next section)[22,23]. The distance between the 3′-overhang-binding pocket of the PAZ domain and the active site of the RNaseIIIa domain is 65 Å, which corresponds to the length of a 25-nucleotide RNA duplex (Fig. 1c). This is in good agreement with the size of siRNA fragments produced by *G. intestinalis* Dicer *in vitro*. The domain architecture of Dicer thus suggests that it functions as a molecular ruler, generating products of defined length by anchoring the 3′ dinucleotide of the substrate RNA duplex (generated by an initial nonspecific cleavage) in the PAZ domain and cleaving at a fixed distance from that end. This is supported by the observation that a truncated *G. intestinalis* Dicer protein that lacks the PAZ domain yields RNA products of variable length *in vitro*[21]. The structure of *G. intestinalis* Dicer suggests that the length of the non-conserved connector helix is the main determinant of product size, providing a possible explanation for the variation in product size across species.

The structure of a fragment of mouse Dicer comprising solely the RNaseIIIb domain and dsRBD was recently solved and indicates that on substrate binding, the dsRBD undergoes a conformational change analogous to that observed in prokaryotic RNaseIII enzymes[24]. In addition, a truncated human Dicer protein lacking the dsRBD has a significantly decreased rate of RNA substrate cleavage *in vitro*, but its substrate affinity remains unaffected[25]. Most Dicer enzymes contain a DEXD/H-box

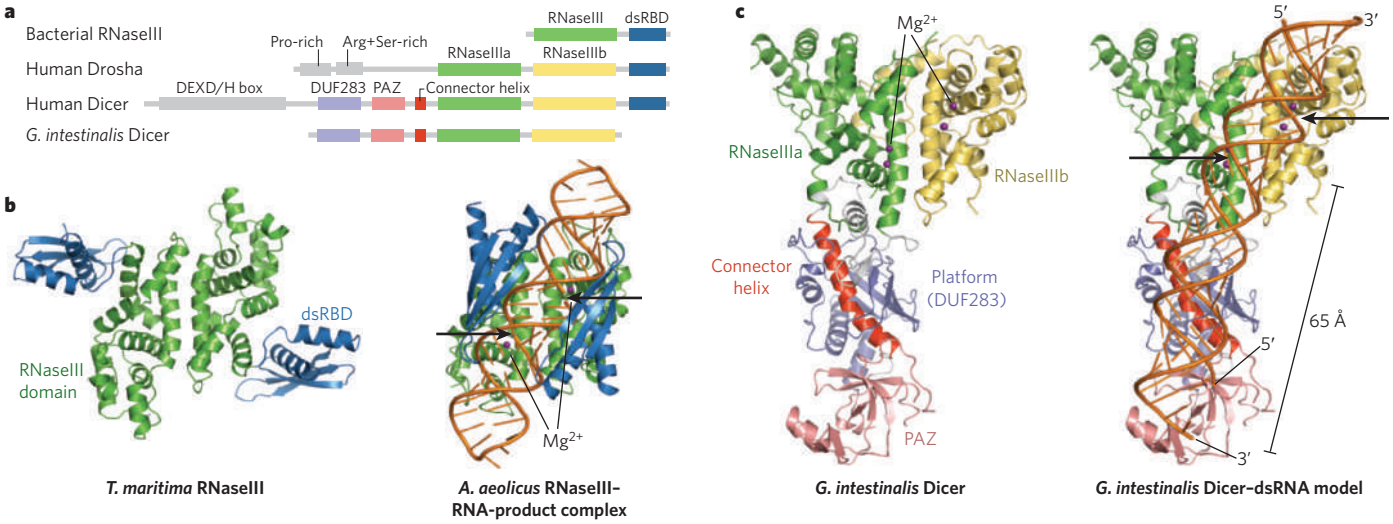


Figure 1 | Structure of RNaseIII-family enzymes. a, A schematic representation of the domain structure of RNaseIII-family enzymes is shown. **b**, The induced fit during dsRNA binding to prokaryotic RNaseIII enzymes, which are homodimeric class I RNaseIII enzymes, is shown. The crystal structure of RNaseIII from the bacterium *Thermotoga maritima* in the RNA-free state (Protein Data Bank (PDB) identity 1O0W) is shown (left). RNaseIII from the bacterium *Aquifex aeolicus* is shown bound to a cleaved dsRNA product (PDB identity 2EZ6) (right). The colours of the protein domains match those in panel **a**; the RNA is shown in gold. Magnesium ions, which are present at the active sites, are shown as purple spheres. Cleavage sites are indicated by arrows. From these structures, it is clear that the dsRNA-binding domains (dsRBDs; blue) undergo a marked rotation on substrate binding. **c**, The molecular mechanism of dsRNA cleavage by *Giardia intestinalis* Dicer is shown. The crystal structure of Dicer (PDB identity 2FFL) is shown (left), together with a model of a dsRNA substrate bound to Dicer (right). In this model, docking of the 3′ overhang of the RNA substrate in the PAZ domain leads to cleavage 65 Å from the 3′ end. Images were generated from files from the PDB using PyMol (http://www.pymol.org).

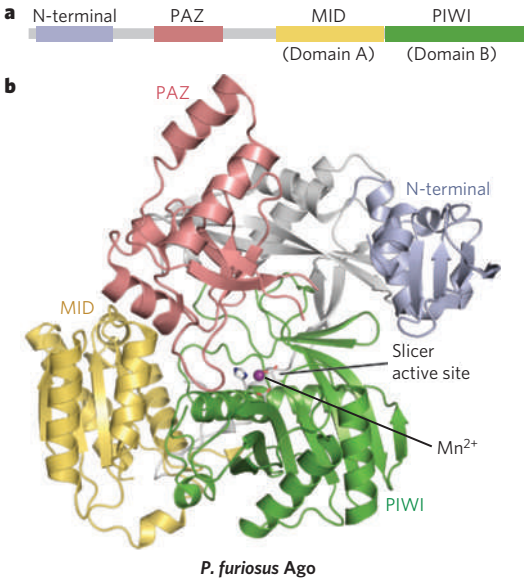


Figure 2 | Modular architecture of Argonaute proteins. a, Eukaryotic Argonaute proteins have four domains: the N-terminal, PAZ, MID and PIWI domains. In some cases, notably for the structures of *Archaeoglobus fulgidus* Piwi protein, the MID domain and PIWI domain have been referred to as domain A and domain B, respectively. **b**, A crystal structure of *Pyrococcus furiosus* Argonaute (Ago) soaked with Mn^{2+} (PDB identity 1Z25) is shown. The protein adopts a bilobate architecture, with the N-terminal and PAZ domains forming one lobe and the MID and PIWI domains forming the other. The metal ion in the active site is shown as a purple sphere. The amino-acid residues involved in metal-ion binding in the slicer catalytic site are shown in stick format.

domain. These domains are found in a diverse group of proteins that are involved in the ATP-dependent binding and remodelling of nucleic acids[26]. Although some invertebrate Dicer proteins (such as *Drosophila melanogaster* DCR-2) seem to require ATP for processive dicing of long dsRNAs (that is, multiple rounds of cleavage without dissociation from the dsRNA substrate), mammalian Dicer proteins seem to be ATP independent[10,25,27,28]. Kinetic analysis of wild-type and mutant human Dicer proteins showed that the DEXD/H-box domain might have an autoinhibitory function, because removal of this domain increases the cleavage rate[25]. This finding suggests that the DEXD/H-box domain imposes on the protein a non-productive conformation that must be rearranged before catalysis can occur. Further structural and biochemical studies of full-length Dicer proteins and their substrate complexes will be necessary to establish whether the DEXD/H-box domain participates in the binding and unwinding of RNA duplexes during the loading of small RNAs into the RISC.

Drosha and the microprocessor complex

The RNaseIII-family member Drosha catalyses the initial processing of pri-miRNAs, yielding pre-miRNAs, which are hairpins with phosphorylated 5′ ends and 3′ dinucleotide overhangs[9]. Drosha is a nuclear protein, and its domain structure consists of a proline-rich region and an arginine- and serine-rich region at the N terminus, followed by two RNaseIII domains and a dsRBD (Fig. 1a). Purified Drosha cleaves dsRNA nonspecifically; specific cleavage of pri-miRNAs requires association with a protein known as DGCR8 (also known as PASHA in invertebrates) in a complex called the microprocessor[29,30]. DGCR8 binds to the base of the pri-miRNA hairpin, positioning Drosha to cleave the pri-miRNA stem at a distance of 11 base pairs from the junction between the duplex stem and the flanking ssRNA regions[31]. Thus, DGCR8 seems to be a *trans*-acting specificity determinant, analogous to the PAZ domain of Dicer, which acts *in cis*. The molecular architecture of Drosha is unknown, but the structure of a core region of human DGCR8, composed of a tandem pair of dsRBDs, was determined recently[32]. The canonical RNA-binding surfaces of the two dsRBDs are non-contiguous, suggesting that the protein binds to two discontinuous segments of the pri-miRNA stem–loop structure.

Structural insights into Argonaute function

The common feature of RNAi and all related small-RNA-mediated silencing pathways is the association of a small silencing RNA (also known as a guide RNA in this context) with a protein of the Argonaute family[4]. The resultant protein–RNA complex forms the minimal core of the effector complex known as the RISC. Within the RISC, the small RNA functions as a sequence-specific guide that recruits an Argonaute protein to complementary target transcripts through base-pairing interactions. The target transcripts, typically mRNAs, are then either cleaved or prevented from being translated by ribosomes, leading to their degradation.

Throughout evolution, the Argonaute family has diverged into specialized clades (or subfamilies) that recognize different small RNA types and confer the specific effects of the various small-RNA silencing pathways[33]. Both siRNAs and miRNAs associate with members of the AGO clade of Argonaute proteins, whereas piRNAs bind to those of the PIWI clade. In classic RNAi, which is elicited by siRNAs, Argonaute proteins silence targeted mRNAs by catalysing their endonucleolytic cleavage, a process known as slicing. The PIWI clade of the Argonaute protein family is thought to use slicing in piRNA-mediated silencing of mobile genetic elements in the germ line[34,35]. During slicing, the target RNA is cleaved at the scissile phosphate group, which is opposite the phosphate group between the tenth and eleventh nucleotides of the guide RNA strand, as measured from the 5′ end of the guide strand[11,12]. Slicing requires perfect complementarity between the guide strand and the target around the cleavage site[36–38]. Argonaute proteins can also silence transcripts independently of slicer activity. In the animal miRNA pathway, Argonaute proteins repress target mRNAs by inhibiting their translation and inducing deadenylation and subsequent mRNA decay. However, the precise mechanism of miRNA-mediated silencing is not fully understood[39].

To function as an effector of small-RNA-mediated silencing, the Argonaute protein must bind to the guide RNA strand, eject the non-guide (passenger) strand of the siRNA or miRNA–miRNA* duplex (where miRNA* is the passenger strand) during loading, and subsequently recognize the target RNA (Box 1). In silencing pathways that rely on RNA slicing, the Argonaute protein carries out multiple cycles of target binding, cleavage and product release, while the guide strand remains bound to the protein. In the metazoan miRNA pathway, in which silencing is achieved in a slicer-independent manner, the Argonaute protein must remain tightly associated with the targeted mRNA to keep its translation repressed.

Functional domains

Argonaute proteins are multidomain proteins that contain an N-terminal domain, and PAZ, middle (MID) and PIWI domains (Fig. 2a). Crystal structures of prokaryotic Argonaute proteins have revealed a bilobate architecture, with the MID and PIWI domains forming one lobe, and the N-terminal and PAZ domains constituting the other (Fig. 2b). The two signature domains of the Argonaute family — the PAZ domain and the C-terminal PIWI domain — were originally identified by phylogenetic sequence analysis[40]. Three-dimensional structures of isolated PAZ domains from *D. melanogaster* Argonaute proteins revealed a fold similar to those of the oligosaccharide/oligonucleotide-binding (OB)-fold domain and Sm-fold domain[17–19]. The first crystal structure of a full-length Argonaute protein, from the archaeal species *Pyrococcus furiosus*, showed that the sequence motif originally defined as the PIWI domain by Lorenzo Cerutti *et al.*[40] consists of two structural domains, termed MID and PIWI, joined by way of an extensive conserved interface that is centred on the buried C terminus of the protein[41] (Fig. 2b). The MID domain resembles the sugar-binding domain of the *lac* repressor. The PIWI domain adopts a fold similar to that in RNaseH, an endoribonuclease that cleaves RNA–DNA hybrids[42]. Crystal structures of a

PIWI-like protein (called Piwi) from the archaeal species *Archaeoglobus fulgidus* (consisting of domain A and domain B, which are equivalent to the MID domain and the PIWI domain, respectively) and of a full-length Argonaute protein from the bacterium *A. aeolicus* also showed the presence of an RNaseH-like fold in their PIWI domains[43,44]. Biochemical studies suggest that, similarly to RNaseH, prokaryotic Argonaute proteins function as DNA-guided ribonucleases[44,45], unlike their eukaryotic counterparts, which have evolved to use RNA molecules as guides.

Slicer activity

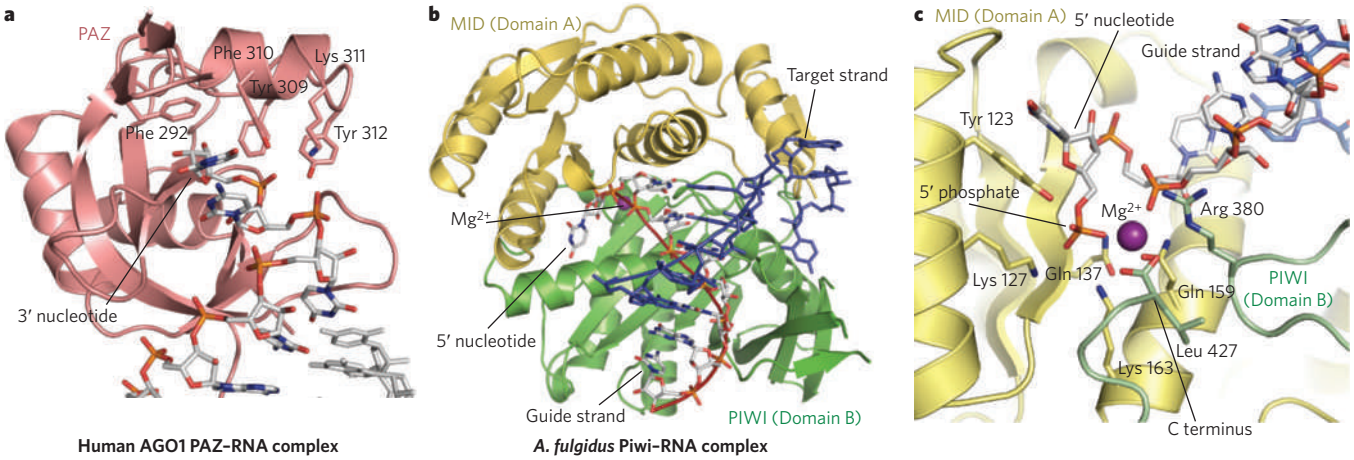
The finding that the PIWI domain of an Argonaute protein adopts an RNaseH fold suggested immediately that Argonaute proteins are responsible for the 'slicer' activity of the RISC[41]. Similarly to the requirements for RNaseH activity, RISC-catalysed RNA cleavage requires divalent metal ions and yields a 5′ product, which has a free 3′ hydroxyl group, and a 3′ product, which carries a 5′ phosphate group[46–48]. The active site of RNaseH contains an Asp-Asp-Glu/Asp motif, which coordinates the two divalent cations required for catalysis[42]. Similarly, in the crystal structure of *P. furiosus* Argonaute (Ago) soaked with manganese ions, a conserved Asp-Asp-His motif was observed to coordinate a single manganese ion[49] (Fig. 2b). Mutagenesis of this motif in human AGO2 confirmed its requirement for slicer activity *in vivo* and *in vitro,* thus establishing the Argonaute protein as the catalytic component of RISCs[46,49]. Similar analysis of *Schizosaccharomyces pombe* (fission yeast) Ago1, a component of the RITS, showed that its slicer activity is required for transcriptional silencing[50]. Of the four human Argonaute proteins (AGO1, AGO2, AGO3 and AGO4), only AGO2 has demonstrable slicer activity[46,51]. The motifs in AGO1 and AGO4 do not precisely conform to the consensus sequence, but AGO3 has a complete Asp-Asp-His sequence and yet is inactive *in vitro*. Moreover, it has recently been shown that *D. melanogaster* PIWI (one of the three members of the PIWI clade of Argonaute proteins in *D. melanogaster*), which has an Asp-Asp-Lys motif, is catalytically active[52], in accordance with the postulated requirement for slicing in the piRNA pathway[34,35]. *D. melanogaster* AGO1 and AGO2 have complete Asp-Asp-His motifs, and both are capable of slicing. But AGO1 is a much less efficient enzyme than AGO2 because it releases products at a slower rate, resulting in a slower turnover[53]. These findings indicate that factors in addition to the conservation of catalytic residues might determine whether a given Argonaute protein is an efficient slicer *in vivo*.

Recognition of RNA termini

The incorporation of siRNAs and miRNAs into the RISC requires the presence of 5′ phosphate groups and 3′ dinucleotide overhangs at the termini[10–12]. The discovery that the PAZ domain is an RNA-binding domain that specifically recognizes the 3′ ends of ssRNAs suggested immediately that it might function as a module for anchoring the 3′ end of the guide RNA strand within the RISC[17–19]. Insights into the mechanism of RNA recognition came from the crystal structure of the PAZ domain of human AGO1 in complex with an siRNA-like duplex[22] and from nuclear magnetic resonance structures of the PAZ domain of *D. melanogaster* AGO2 in complex with ssRNA oligonucleotides[23]. In these structures, the 3′ dinucleotide inserts into a preformed hydrophobic pocket that is lined with conserved aromatic residues (Fig. 3a). Although there are no sequence-specific contacts, the base of the terminal nucleotide stacks against the aromatic ring of a conserved phenylalanine residue (Fig. 3a).

The 5′ phosphate groups of siRNAs and miRNAs, which result from their mechanism of biogenesis, are crucial for the efficient assembly of these small RNAs into the RISC[10,12,38,46]. Moreover, the 5′ phosphate group is essential for slicing fidelity, because the position of the cleavage site in the target RNA strand is determined by its distance from the 5′ phosphate group of the guide RNA strand[12,49,54]. Crystal structures of the *A. fulgidus* Argonaute protein Piwi bound to a short dsRNA duplex (which mimics the interaction between the guide strand and the target strand within the RISC) showed that the 5′ nucleotide of the guide strand is distorted and does not base-pair with the target strand of the RNA duplex[45,55] (Fig. 3b). This is consistent with the observation that a base mismatch at this position is tolerated and can increase slicer activity[56]. The 5′ phosphate group of the guide RNA strand is buried in a deep pocket at the interface between the MID domain and the PIWI domain and is bound to a magnesium ion that is, in turn, coordinated to the protein's C terminus (Fig. 3c). Mutation of any of the four residues involved in metal-ion coordination and 5′-phosphate binding, which are among the most highly conserved residues in the Argonaute protein family, impairs slicing activity[45]. This underscores the functional importance of the 5′-phosphate-binding pocket in anchoring the guide RNA.

In the plant *Arabidopsis thaliana*, distinct types of small non-coding RNA associate with different Argonaute proteins on the basis of the identity of the 5′ nucleotide[57,58]. *A. thaliana* AGO1 binds mainly to RNAs with a uridine at their 5′ end, whereas AGO2 recruits RNAs with



Human AGO1 PAZ–RNA complex

A. fulgidus Piwi–RNA complex

Figure 3| Recognition of the termini of small RNAs by the PAZ and MID domains of Argonaute proteins. a, The crystal structure of the PAZ domain of human AGO1 (ribbon format) is shown in complex with an siRNA-like duplex (stick format) (PDB identity 1SI3). Conserved residues in contact with the 3′ nucleotide are shown in stick format and labelled. The 3′ end of the siRNA inserts into a preformed hydrophobic pocket, with the base of the 3′ nucleotide stacking against an invariant aromatic residue (the phenylalanine at position 292 in the protein, Phe 292). **b**, The crystal structure of *Archaeoglobus fulgidus* Piwi bound to an RNA duplex (PDB identity 2BGG), which mimics the guide–target interaction, is shown. *A. fulgidus* Piwi is composed of a MID domain and a PIWI domain (also referred to as domain A and domain B, respectively). The 5′ nucleotide of the guide strand (backbone shown in red) binds at the MID–PIWI domain interface and does not base-pair with the target strand (blue). **c**, Shown is a detailed view of the pocket in *A. fulgidus* Piwi that binds to the 5′ end of small RNAs. A magnesium ion, coordinated by the C terminus of the protein (Leu 427) and the 5′ phosphate group, is shown as a purple sphere. Conserved residues that are involved in metal-ion coordination and 5′-phosphate binding are shown in stick format and labelled.

a 5′ adenosine. Domain-swap experiments involving chimaeric proteins showed that the specificity of the 5′ nucleotide is conferred by the MID and PIWI domains of the Argonaute proteins[57]. Similarly, the PIWI clade of the Argonaute protein family associates with piRNAs, which are ~24–31-nucleotide RNA species that are 2′-O-methylated and have a bias for uridine as the first nucleotide[59,60]. It is thus probable that the PAZ and MID domains of Argonaute proteins that bind to different small RNA species underwent an evolutionary adaptation that allows small RNAs to be sorted on the basis of particular 3′ and 5′ ends.

RISC loading

In the siRNA- and miRNA-mediated pathways of gene silencing, the loading of a guide RNA onto an Argonaute protein is closely linked with its biogenesis (Box 1). Starting from an siRNA duplex or a miRNA–miRNA* generated by Dicer-mediated cleavage, the strand with its 5′ end at the less thermodynamically stable end of the duplex is selected as the guide strand[61]. The observation, made from the structure of *A. fulgidus* Piwi bound to an siRNA-like duplex, that the first nucleotide of the guide strand is unpaired and buried in the 5′-phosphate-binding pocket suggests that this mode of guide-strand binding contributes to the apparent guide-strand selectivity[45,55]. Argonaute proteins also promote guide-strand selection by subsequently slicing the passenger strand, thereby facilitating its release[62–65]. However, for duplexes with multiple mismatches, as is often the case in the miRNA pathway, slicing is not required during loading[63]. Thus, Argonaute proteins lacking slicer activity (such as human AGO1, AGO3 and AGO4) can still be loaded with miRNAs.

RISC loading takes place in the context of the RISC-loading complex. In human cells, the RISC-loading complex consists of an Argonaute protein, Dicer and the dsRBD-containing protein TRPB[66–68]. *In vitro,* this ternary complex can process a pre-miRNA, load the correct guide strand and cleave a target RNA, all in the absence of ATP[67]. What are the roles of dsRBD-containing proteins in substrate selection and RISC loading? In *D. melanogaster*, DCR-1 associates with the dsRBD-containing protein Loquacious (LOQS) to convert pre-miRNAs into miRNA–miRNA* duplexes[69–71]. By contrast, a complex of DCR-2 with its dsRBD-containing protein partner, R2D2, is required for the efficient production of siRNAs from long dsRNA precursors and for their subsequent loading onto AGO2 (refs 63, 72–74). Recent studies indicate that the structure and the extent of base-pairing (that is, the presence of mismatches) within siRNA and miRNA precursor duplexes determine whether the guide strands associate with AGO1 or AGO2 in *D. melanogaster*[53,75]. The affinity of DCR-2–R2D2 is higher for perfectly matched duplexes than for 'bulged' duplexes (such as pre-miRNA hairpins), and this seems to be the main source of small RNA selectivity during AGO2 loading[75].

Target recognition

In the *A. fulgidus* Piwi–RNA complex structure, the guide–target duplex is positioned over a conserved basic channel that spans the surfaces of both the MID domain and the PIWI domain[45,55] (Fig. 3c). Importantly, these structures show that the bases of nucleotides 2–6 of the guide strand (starting from the 5′ end) are exposed and free to base-pair with a target mRNA. This agrees well with numerous computational and biochemical studies indicating that nucleotides 2–8 of the guide strand (known as the 'seed' region) are crucial determinants of the specificity of target recognition by miRNAs[76–78]. Modelling the trajectory of a full-length siRNA–target duplex places the scissile phosphate group of the target RNA in the putative slicer catalytic site, thus providing a rationale for the specificity of slicer cleavage at a fixed distance from the 5′ end of the guide strand[11,12]. Perfect complementarity around the cleavage site in the guide–target duplex is a prerequisite for slicing[36–38]. Presumably, base-pairing around the 10–11-base step ensures correct orientation of the scissile phosphate group in the active site.

In the context of the bilobate architecture of full-length Argonaute proteins, the guide–target duplex has been proposed to bind in a positively charged cleft between the PAZ–N-terminal and MID–PIWI lobes[41,44,49]. Recent crystal structures of full-length *Thermus thermophilus* Argonaute protein bound to guide DNA strands have identified the molecular details of guide-strand recognition[79]. The structure of *T. thermophilus* Ago in complex with a 5′-phosphorylated 21-nucleotide DNA shows the guide strand bound with its 5′ phosphate group in the MID domain and with its 3′ end in the PAZ domain[79] (Fig. 4a). Through interactions with conserved arginine residues, nucleotides 2–10 of the guide strand adopt a stacked helical conformation. Thus, the seed region of the guide strand is pre-organized to initiate base-pairing with the target strand (Fig. 4a). In the absence of a target strand, the guide strand is kinked at the 10–11-base step, indicating that a further structural rearrangement occurs on target-RNA recognition.

The most recent insights into the mechanism of target-RNA recognition come from the crystal structure of a ternary complex consisting of *T. thermophilus* Ago, a 21-nucleotide guide DNA strand and a 20-nucleotide target RNA with mismatched bases introduced

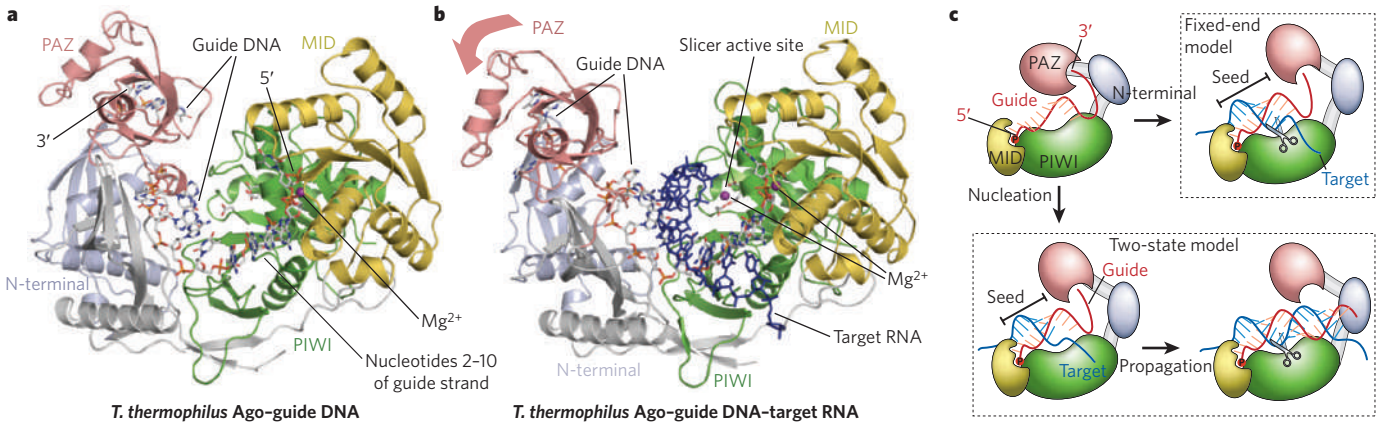


Figure 4 | Mechanism of guide- and target-strand recognition by Argonaute proteins. a, The crystal structure of *Thermus thermophilus* Ago (ribbon format) bound to a 5′-phosphorylated 21-nucleotide guide DNA strand (stick format) (PDB identity 3DLH) is shown. Nucleotides 1–11 and 18–21 of the guide DNA are visible in the structure and shown in stick format. The 5′ phosphate group of the guide strand binds to the MID domain, whereas the 3′ end is anchored in the PAZ domain. **b**, The crystal structure of a ternary complex of *T. thermophilus* Ago with a 5′-phosphorylated 21-nucleotide guide DNA strand and a 20-nucleotide target RNA (PDB identity 3F73) is shown. The complex is shown in the same orientation as the binary complex in panel **a**, after superposition of the PIWI domains. The target RNA (blue) and guide DNA (grey backbone) are shown in stick format. The arrow indicates the conformational change undergone by the lobe containing the PAZ and N-terminal domains on target-RNA binding. **c**, Putative models for target-RNA recognition by Argonaute proteins are illustrated. The fixed-end model postulates that both the 5′ and 3′ ends of the guide strand remain docked in their binding pockets during slicing. The two-state model postulates that the seed region of the guide occurs in two steps: first, nucleation of base-pairing with the target RNA; and second, propagation of the guide–target duplex, leading to the release of the 3′ end of the guide strand from the PAZ domain.

at the 10–11-base step[80] (Fig. 4b). Both termini of the guide strand are anchored in their respective binding pockets, as is the case for the *T. thermophilus* Ago–guide DNA complex. The seed region of the guide strand (nucleotides 2–8) engages in Watson–Crick base-pairing interactions with the target RNA, assuming an A-form helical conformation. To accommodate the target RNA strand in the central channel, Ago undergoes a pronounced conformational change to a more open conformation, achieved by rotating the N-terminal and PAZ domains away from the lobe containing the MID and PIWI domains.

Two models for target recognition and cleavage by Argonaute proteins have been proposed[81,82]. The fixed-end model postulates that both ends of the guide RNA remain bound to the Argonaute protein during slicing. This presents a topological constraint on the guide–target interaction that would limit the extent of base-pairing to less than one helical turn (11 base pairs) and might be the factor that limits the seed region to nucleotides 2–8 of the guide RNA. By contrast, the two-state model proposes that the target binds to the seed region of the guide strand, and then propagation of base-pairing towards the 3′ end of the guide strand results in this end of the guide strand dissociating from the PAZ domain. At present, it is unclear whether the structure of the *T. thermophilus* Ago ternary complex represents a cleavage-competent complex as envisaged by the fixed-end model or whether it depicts a trapped intermediate as postulated by the two-state model. Further structural studies of catalytically inactive Argonaute proteins in complexes with perfectly matched guide–target duplexes will be necessary to clarify this issue. Nonetheless, the mode of target-RNA recognition observed in the *T. thermophilus* Ago ternary complex might be representative of target mRNA recognition during miRNA-mediated silencing in metazoans, in which slicing is typically prevented by a base mismatch between the miRNA and the target at around the 10–11-base step.

Slicer-independent functions

In metazoans, gene silencing by miRNAs occurs by the mere anchoring of a miRNA-containing RISC to the target mRNA, in the absence of slicer-dependent cleavage of the target mRNA[83]. The mechanism of miRNA-mediated repression is unclear, but it must involve interactions between the RISC and the cellular machineries that are responsible for translation and mRNA decay. Recent studies using *in vitro* systems indicate that miRNAs might affect the initiation of translation by interfering with the function of the EIF4F, a complex that binds to the mRNA cap (7-methylguanosine)[84]. The MID domain of human AGO2 shows limited sequence homology to the cap-binding motif of the translation initiation factor EIF4E, a subunit of EIF4F, suggesting that human AGO2 competes with EIF4E for binding to the mRNA cap structure[85]. However, the MID domain of Argonaute proteins lacks any discernible structural homology to EIF4E. Thus, the AGO2–cap interaction, if indeed direct, might be a non-canonical mode of mRNA cap recognition.

Argonaute proteins, miRNAs and their mRNA targets co-localize in P bodies, which are cytoplasmic foci where mRNA decay is thought to occur. In mammalian cells and in *D. melanogaster*, the interaction between Argonaute proteins and the P-body protein GW182 is required for miRNA-mediated repression of mRNAs, as well as for Argonaute localization[86–89]. It was shown recently that the glycine- and tryptophan-rich (GW) motifs of GW182 directly interact with the MID–PIWI region of human Argonaute proteins and that a minimal fragment (termed the Ago hook), encompassing two tandem GW motifs, is sufficient to mediate this interaction[90]. Interestingly, mutations that disrupt the Argonaute–GW182 interaction have been mapped mainly to the 5′-phosphate-binding pocket located at the interface between the MID domain and PIWI domain[89,90].

Future prospects

Structural studies and structure-based biochemical studies will continue to improve our understanding of the mechanisms and evolutionary relationships of RNAi pathways. An important future goal will be to determine the structures of some of the proteins and protein complexes that operate in eukaryotic cells. Elucidating the molecular architecture of the RISC-loading complex will bring important insights into the coupling of small RNA biogenesis, Argonaute loading and guide-strand selection. Recent proteomic studies have uncovered a multitude of Argonaute-interacting proteins (for example GW182, MOV10 and FXR1) that might link Argonaute proteins to downstream effector events[91,92]. Obtaining structural and structure-based functional insights into these interactions will help to uncover the molecular mechanisms that underlie the functions of Argonaute proteins in small-RNA-mediated gene silencing. Results from these studies will not only provide new mechanistic details but also lay the foundation for the informed engineering of RNAi as a therapeutic tool. ■

1. Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120,** 15–20 (2005).
2. Stefani, G. & Slack, F. J. Small non-coding RNAs in animal development. *Nature Rev. Mol. Cell Biol.* **9,** 219–230 (2008).
3. Bernstein, E., Caudy, A. A., Hammond, S. M. & Hannon, G. J. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* **409,** 363–366 (2001).
4. Hutvagner, G. & Simard, M. J. Argonaute proteins: key players in RNA silencing. *Nature Rev. Mol. Cell Biol.* **9,** 22–32 (2008).
5. Grewal, S. I. & Elgin, S. C. Transcription and RNA interference in the formation of heterochromatin. *Nature* **447,** 399–406 (2007).
6. Klattenhoff, C. & Theurkauf, W. Biogenesis and germline functions of piRNAs. *Development* **135,** 3–9 (2008).
7. Haasnoot, J., Westerhout, E. M. & Berkhout, B. RNA interference against viruses: strike and counterstrike. *Nature Biotechnol.* **25,** 1435–1443 (2007).
8. Li, F. & Ding, S. W. Virus counterdefense: diverse strategies for evading the RNA-silencing immunity. *Annu. Rev. Microbiol.* **60,** 503–531 (2006).
9. Lee, Y. *et al.* The nuclear RNase III Drosha initiates microRNA processing. *Nature* **425,** 415–419 (2003).
10. Nykänen, A., Haley, B. & Zamore, P. D. ATP requirements and small interfering RNA structure in the RNA interference pathway. *Cell* **107,** 309–321 (2001).
11. Elbashir, S. M., Lendeckel, W. & Tuschl, T. RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev.* **15,** 188–200 (2001).
12. Elbashir, S. M., Martinez, J., Patkaniowska, A., Lendeckel, W. & Tuschl, T. Functional anatomy of siRNAs for mediating efficient RNAi in *Drosophila melanogaster* embryo lysate. *EMBO J.* **20,** 6877–6888 (2001).
13. Macrae, I. J. & Doudna, J. A. Ribonuclease revisited: structural insights into ribonuclease III family enzymes. *Curr. Opin. Struct. Biol.* **17,** 138–145 (2007).
14. Blaszczyk, J. *et al.* Crystallographic and modeling studies of RNase III suggest a mechanism for double-stranded RNA cleavage. *Structure* **9,** 1225–1236 (2001).
15. Zhang, H., Kolb, F. A., Jaskiewicz, L., Westhof, E. & Filipowicz, W. Single processing center models for human Dicer and bacterial RNase III. *Cell* **118,** 57–68 (2004).
This paper elegantly showed that Dicer contains a single processing centre and proposed that the two RNaseIII domains of Dicer function as an intramolecular dimer.
16. Gan, J. *et al.* Structural insight into the mechanism of double-stranded RNA processing by ribonuclease III. *Cell* **124,** 355–366 (2006).
17. Lingel, A., Simon, B., Izaurralde, E. & Sattler, M. Structure and nucleic-acid binding of the *Drosophila* Argonaute 2 PAZ domain. *Nature* **426,** 465–469 (2003).
18. Yan, K. S. *et al.* Structure and conserved RNA binding of the PAZ domain. *Nature* **426,** 468–474 (2003).
19. Song, J. J. *et al.* The crystal structure of the Argonaute2 PAZ domain reveals an RNA binding motif in RNAi effector complexes. *Nature Struct. Biol.* **10,** 1026–1032 (2003).
20. Macrae, I. J. *et al.* Structural basis for double-stranded RNA processing by Dicer. *Science* **311,** 195–198 (2006).
This paper described the three-dimensional architecture of Dicer, revealing how Dicer functions as a molecular ruler to determine the length of its dsRNA products.
21. Macrae, I. J., Zhou, K. & Doudna, J. A. Structural determinants of RNA recognition and cleavage by Dicer. *Nature Struct. Mol. Biol.* **14,** 934–940 (2007).
22. Ma, J. B., Ye, K. & Patel, D. J. Structural basis for overhang-specific small interfering RNA recognition by the PAZ domain. *Nature* **429,** 318–322 (2004).
This paper revealed the mode of recognition of the 3′ end of siRNAs by the PAZ domain, identifying conserved residues that bind the 3′ nucleotide.
23. Lingel, A., Simon, B., Izaurralde, E. & Sattler, M. Nucleic acid 3′-end recognition by the Argonaute2 PAZ domain. *Nature Struct. Mol. Biol.* **11,** 576–577 (2004).
This paper showed that the PAZ domain provides a conserved hydrophobic binding pocket for the 3′ nucleotide of ssRNA.
24. Du, Z., Lee, J. K., Tjhen, R., Stroud, R. M. & James, T. L. Structural and biochemical insights into the dicing mechanism of mouse Dicer: a conserved lysine is critical for dsRNA cleavage. *Proc. Natl Acad. Sci. USA* **105,** 2391–2396 (2008).
25. Ma, E., Macrae, I. J., Kirsch, J. F. & Doudna, J. A. Autoinhibition of human Dicer by its internal helicase domain. *J. Mol. Biol.* **380,** 237–243 (2008).
26. Jankowsky, E. & Fairman, M. E. RNA helicases — one fold for many functions. *Curr. Opin. Struct. Biol.* **17,** 316–324 (2007).
27. Provost, P. *et al.* Ribonuclease activity and RNA binding of recombinant human Dicer. *EMBO J.* **21,** 5864–5874 (2002).
28. Zhang, H., Kolb, F. A., Brondani, V., Billy, E. & Filipowicz, W. Human Dicer preferentially cleaves dsRNAs at their termini without a requirement for ATP. *EMBO J.* **21,** 5875–5885 (2002).
29. Han, J. *et al.* The Drosha–DGCR8 complex in primary microRNA processing. *Genes Dev.* **18,** 3016–3027 (2004).
30. Gregory, R. I. *et al.* The Microprocessor complex mediates the genesis of microRNAs. *Nature* **432,** 235–240 (2004).

31. Han, J. *et al.* Molecular basis for the recognition of primary microRNAs by the Drosha–DGCR8 complex. *Cell* **125,** 887–901 (2006).
32. Sohn, S. Y. *et al.* Crystal structure of human DGCR8 core. *Nature Struct. Mol. Biol.* **14,** 847–853 (2007).
33. Farazi, T. A., Juranek, S. A. & Tuschl, T. The growing catalog of small RNAs and their association with distinct Argonaute/Piwi family members. *Development* **135,** 1201–1214 (2008).
34. Brennecke, J. *et al.* Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell* **128,** 1089–1103 (2007).
35. Gunawardane, L. S. *et al.* A slicer-mediated mechanism for repeat-associated siRNA 5′ end formation in *Drosophila*. *Science* **315,** 1587–1590 (2007).
36. Hutvágner, G. & Zamore, P. D. A microRNA in a multiple-turnover RNAi enzyme complex. Science **297,** 2056–2060 (2002).
37. Doench, J. G., Petersen, C. P. & Sharp, P. A. siRNAs can function as miRNAs. *Genes Dev.* **17,** 438–442 (2003).
38. Chiu, Y. L. & Rana, T. M. RNAi in human cells: basic structural and functional features of small interfering RNA. *Mol. Cell* **10,** 549–561 (2002).
39. Jackson, R. J. & Standart, N. How do microRNAs regulate gene expression? *Sci. STKE* **2007,** re1 (2007).
40. Cerutti, L., Mian, N. & Bateman, A. Domains in gene silencing and cell differentiation proteins: the novel PAZ domain and redefinition of the Piwi domain. *Trends Biochem. Sci.* **25,** 481–482 (2000).
41. Song, J. J., Smith, S. K., Hannon, G. J. & Joshua-Tor, L. Crystal structure of Argonaute and its implications for RISC slicer activity. *Science* **305,** 1434–1437 (2004).
This paper revealed the molecular architecture of Argonaute proteins and showed that the PIWI domain resembles RNaseH, suggesting that Argonaute is the 'slicer'.
42. Nowotny, M., Gaidamakov, S. A., Crouch, R. J. & Yang, W. Crystal structures of RNase H bound to an RNA/DNA hybrid: substrate specificity and metal-dependent catalysis. *Cell* **121,** 1005–1016 (2005).
43. Parker, J. S., Roe, S. M. & Barford, D. Crystal structure of a PIWI protein suggests mechanisms for siRNA recognition and slicer activity. *EMBO J.* **23,** 4727–4237 (2004).
44. Yuan, Y. R. *et al.* Crystal structure of *A. aeolicus* argonaute, a site-specific DNA-guided endoribonuclease, provides insights into RISC-mediated mRNA cleavage. *Mol. Cell* **19,** 405–419 (2005).
45. Ma, J. B. *et al.* Structural basis for 5′-end-specific recognition of guide RNA by the *A. fulgidus* Piwi protein. *Nature* **434,** 666–670 (2005).
This paper revealed that the 5′ phosphate group of the guide RNA strand binds to a pocket at the interface of the MID and PIWI domains in the *A. fulgidus* Piwi protein and that the first nucleotide of the guide strand does not base-pair with the target RNA.
46. Liu, J. *et al.* Argonaute2 is the catalytic engine of mammalian RNAi. *Science* **305,** 1437–1441 (2004).
This paper showed that human AGO2 has slicer activity and that mutations in its PIWI domain, based on the structure of archaeal Argonaute, abolish RISC activity *in vivo*.
47. Martinez, J. & Tuschl, T. RISC is a 5′ phosphomonoester-producing RNA endonuclease. *Genes Dev.* **18,** 975–980 (2004).
48. Schwarz, D. S., Tomari, Y. & Zamore, P. D. The RNA-induced silencing complex is a Mg^{2+}-dependent endonuclease. *Curr. Biol.* **14,** 787–791 (2004).
49. Rivas, F. V. *et al.* Purified Argonaute2 and an siRNA form recombinant human RISC. *Nature Struct. Mol. Biol.* **12,** 340–349 (2005).
50. Irvine, D. V. *et al.* Argonaute slicing is required for heterochromatic silencing and spreading. *Science* **313,** 1134–1137 (2006).
51. Meister, G. *et al.* Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. *Mol. Cell* **15,** 185–197 (2004).
52. Saito, K. *et al.* Specific association of Piwi with rasiRNAs derived from retrotransposon and heterochromatic regions in the *Drosophila* genome. *Genes Dev.* **20,** 2214–2222 (2006).
53. Förstemann, K., Horwich, M. D., Wee, L., Tomari, Y. & Zamore, P. D. *Drosophila* microRNAs are sorted into functionally distinct Argonaute complexes after production by Dicer-1. *Cell* **130,** 287–297 (2007).
54. Elbashir, S. M. *et al.* Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature* **411,** 494–498 (2001).
55. Parker, J. S., Roe, S. M. & Barford, D. Structural insights into mRNA recognition from a PIWI domain–siRNA guide complex. *Nature* **434,** 663–666 (2005).
The paper revealed how the 5′ phosphate group of the guide RNA strand is recognized by Argonaute, and it highlights the importance of the seed region in mediating guide–target recognition.
56. Haley, B. & Zamore, P. D. Kinetic analysis of the RNAi enzyme complex. *Nature Struct. Mol. Biol.* **11,** 599–606 (2004).
57. Mi, S. *et al.* Sorting of small RNAs into *Arabidopsis* argonaute complexes is directed by the 5′ terminal nucleotide. *Cell* **133,** 116–127 (2008).
58. Montgomery, T. A. *et al.* Specificity of ARGONAUTE7–miR390 interaction and dual functionality in *TAS3 trans*-acting siRNA formation. *Cell* **133,** 128–141 (2008).
59. Ohara, T. *et al.* The 3′ termini of mouse Piwi-interacting RNAs are 2′-O-methylated. *Nature Struct. Mol. Biol.* **14,** 349–350 (2007).
60. Kirino, Y. & Mourelatos, Z. Mouse Piwi-interacting RNAs are 2′-O-methylated at their 3′ termini. *Nature Struct. Mol. Biol.* **14,** 347–348 (2007).
61. Schwarz, D. S. *et al.* Asymmetry in the assembly of the RNAi enzyme complex. *Cell* **115,** 199–208 (2003).
62. Miyoshi, K., Tsukumo, H., Nagami, T., Siomi, H. & Siomi, M. C. Slicer function of *Drosophila* Argonautes and its involvement in RISC formation. *Genes Dev.* **19,** 2837–2848 (2005).
63. Matranga, C., Tomari, Y., Shin, C., Bartel, D. P. & Zamore, P. D. Passenger-strand cleavage facilitates assembly of siRNA into Ago2-containing RNAi enzyme complexes. *Cell* **123,** 607–620 (2005).
64. Rand, T. A., Petersen, S., Du, F. & Wang, X. Argonaute2 cleaves the anti-guide strand of siRNA during RISC activation. *Cell* **123,** 621–629 (2005).
65. Leuschner, P. J., Ameres, S. L., Kueng, S. & Martinez, J. Cleavage of the siRNA passenger strand during RISC assembly in human cells. *EMBO Rep.* **7,** 314–320 (2006).
66. Gregory, R. I., Chendrimada, T. P., Cooch, N. & Shiekhattar, R. Human RISC couples microRNA biogenesis and posttranscriptional gene silencing. *Cell* **123,** 631–640 (2005).
67. Macrae, I. J., Ma, E., Zhou, M., Robinson, C. V. & Doudna, J. A. *In vitro* reconstitution of the human RISC-loading complex. *Proc. Natl Acad. Sci. USA* **105,** 512–517 (2008).
68. Maniataki, E. & Mourelatos, Z. A human, ATP-independent, RISC assembly machine fueled by pre-miRNA. *Genes Dev.* **19,** 2979–2990 (2005).
69. Förstemann, K. *et al.* Normal microRNA maturation and germ-line stem cell maintenance requires Loquacious, a double-stranded RNA-binding domain protein. *PLoS Biol.* **3,** e236 (2005).
70. Saito, K., Ishizuka, A., Siomi, H. & Siomi, M. C. Processing of pre-microRNAs by the Dicer-1–Loquacious complex in *Drosophila* cells. *PLoS Biol.* **3,** e235 (2005).
71. Jiang, F. *et al.* Dicer-1 and R3D1-L catalyze microRNA maturation in *Drosophila*. *Genes Dev.* **19,** 1674–1679 (2005).
72. Tomari, Y. *et al.* RISC assembly defects in the *Drosophila* RNAi mutant *armitage*. *Cell* **116,** 831–841 (2004).
73. Liu, X., Jiang, F., Kalidas, S., Smith, D. & Liu, Q. Dicer-2 and R2D2 coordinately bind siRNA to promote assembly of the siRISC complexes. *RNA* **12,** 1514–1520 (2006).
74. Liu, Q. *et al.* R2D2, a bridge between the initiation and effector steps of the *Drosophila* RNAi pathway. *Science* **301,** 1921–1925 (2003).
75. Tomari, Y., Du, T. & Zamore, P. D. Sorting of *Drosophila* small silencing RNAs. *Cell* **130,** 299–308 (2007).
76. Stark, A., Brennecke, J., Russell, R. B. & Cohen, S. M. Identification of *Drosophila* microRNA targets. *PLoS Biol.* **1,** e60 (2003).
77. Lewis, B. P., Shih, I. H., Jones-Rhoades, M. W., Bartel, D. P. & Burge, C. B. Prediction of mammalian microRNA targets. *Cell* **115,** 787–798 (2003).
78. Doench, J. G. & Sharp, P. A. Specificity of microRNA target selection in translational repression. *Genes Dev.* **18,** 504–511 (2004).
79. Wang, Y., Sheng, G., Juranek, S. A., Tuschl, T. & Patel, D.J. Structure of the guide-strand-containing argonaute silencing complex. *Nature* **456,** 209–213 (2008).
This paper revealed that binding of the guide strand to Argonaute orders the seed region in a helical conformation, poised to initiate base-pairing with the target strand.
80. Wang, Y. *et al.* Structure of an argonaute silencing complex with a seed-containing guide DNA and target RNA duplex. *Nature* **456,** 921–926 (2008).
81. Filipowicz, W. RNAi: the nuts and bolts of the RISC machine. *Cell* **122,** 17–20 (2005).
82. Tomari, Y. & Zamore, P. D. Machines for RNAi. *Genes Dev.* **19,** 517–529 (2005).
83. Pillai, R. S., Artus, C. G. & Filipowicz, W. Tethering of human Ago proteins to mRNA mimics the miRNA-mediated repression of protein synthesis. *RNA* **10,** 1518–1525 (2004).
84. Mathonnet, G. *et al.* MicroRNA inhibition of translation initiation *in vitro* by targeting the cap-binding complex eIF4F. *Science* **317,** 1764–1767 (2007).
85. Kiriakidou, M. *et al.* An mRNA m^7G cap binding-like motif within human Ago2 represses translation. *Cell* **129,** 1141–1151 (2007).
86. Jakymiw, A. *et al.* Disruption of GW bodies impairs mammalian RNA interference. *Nature Cell Biol.* **7,** 1267–1274 (2005).
87. Behm-Ansmant, I. *et al.* mRNA degradation by miRNAs and GW182 requires both CCR4:NOT deadenylase and DCP1:DCP2 decapping complexes. *Genes Dev.* **20,** 1885–1898 (2006).
88. Liu, J. *et al.* A role for the P-body component GW182 in microRNA function. *Nature Cell Biol.* **7,** 1261–1266 (2005).
89. Eulalio, A., Huntzinger, E. & Izaurralde, E. GW182 interaction with Argonaute is essential for miRNA-mediated translational repression and mRNA decay. *Nature Struct. Mol. Biol.* **15,** 346–353 (2008).
90. Till, S. *et al.* A conserved motif in Argonaute-interacting proteins mediates functional interactions through the Argonaute PIWI domain. *Nature Struct. Mol. Biol.* **14,** 897–903 (2007).
91. Meister, G. *et al.* Identification of novel argonaute-associated proteins. *Curr. Biol.* **15,** 2149–2155 (2005).
92. Höck, J. *et al.* Proteomic and functional analysis of Argonaute-containing mRNA–protein complexes in human cells. *EMBO Rep.* **8,** 1052–1060 (2007).

Acknowledgements We are grateful to D. Patel for communicating results in advance of publication. We also thank members of the Doudna laboratory for discussions and critical reading of the manuscript. Research in the Doudna laboratory is supported by the Howard Hughes Medical Institute and the National Institutes of Health. M.J. was supported by the European Molecular Biology Organization and is now a postdoctoral fellow of the Human Frontier Science Program.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence should be addressed to J.A.D. (doudna@berkeley.edu).

Small RNAs in transcriptional gene silencing and genome defence

Danesh Moazed[1]

Small RNA molecules of about 20–30 nucleotides have emerged as powerful regulators of gene expression and genome stability. Studies in fission yeast and multicellular organisms suggest that effector complexes, directed by small RNAs, target nascent chromatin-bound non-coding RNAs and recruit chromatin-modifying complexes. Interactions between small RNAs and nascent non-coding transcripts thus reveal a new mechanism for targeting chromatin-modifying complexes to specific chromosome regions and suggest possibilities for how the resultant chromatin states may be inherited during the process of chromosome duplication.

RNA interference (RNAi) originally referred to the ability of exogenously introduced double-stranded RNA (dsRNA) molecules to silence the expression of homologous sequences in the nematode *Caenorhabditis elegans*[1]. It has become clear over the past decade that RNAi is mechanistically related to a number of other conserved RNA silencing pathways, which are involved in the cellular control of gene expression and in protection of the genome against mobile repetitive DNA sequences, retroelements and transposons[2–4]. These RNA silencing pathways are all associated with small (~20–30 nucleotide) RNAs that function as specificity factors for inactivating homologous sequences by a variety of mechanisms. At least three classes of small RNA have been identified so far (Table 1). The first two classes, short interfering RNAs (siRNAs) and microRNAs (miRNAs), are ~21–25 nucleotides and are generated from longer dsRNA precursors by Dicer, a ribonuclease III (RNaseIII) enzyme. They are loaded into the RNA-induced silencing complex (RISC) or a nuclear form of RISC, called the RNA-induced transcriptional silencing complex (RITS)[5–10]. RISC and RITS are effector complexes that are targeted to homologous sequences by base-pairing interactions involving the guide strand of the small RNA. The core component of each complex is a highly conserved PAZ- and PIWI-domain-containing protein called Argonaute, which binds to the guide small RNA by means of interactions that involve its PAZ domain, as well as the PIWI and middle (MID) domains, and cleaves the target RNA by means of its RNaseH-like PIWI domain (see page 405 for further information about the structural biology of RNAi proteins).

The Argonaute family of proteins, together with the small RNAs that program them, are the central players in RNA silencing, and seem to participate in all small-RNA silencing pathways thus far described. Phylogenetically, Argonaute-family proteins are divided into the AGO and PIWI clades[11]. The PIWI-clade proteins bind to a third class of small RNAs, called PIWI-interacting RNAs (piRNAs), which have a broader average size (~24–31 nucleotides) than siRNAs and miRNAs and are involved in defence against parasitic DNA elements[12–18]. As discussed later, piRNA-programmed PIWI-clade proteins are also likely to function as RISC- and RITS-like complexes that target the inactivation of homologous sequences (Table 1). With the notable exception of budding yeast, small-RNA-mediated silencing mechanisms and their role in chromatin regulation are conserved throughout eukaryotes, indicating an ancient evolutionary origin.

This Review discusses the roles of diverse small-RNA silencing pathways in the regulation of chromatin structure and transcription in plants, animals and fungi, with particular emphasis on emerging common themes. In addition to their well-known roles in post-transcriptional gene silencing (PTGS), in which silencing is directed at the level of messenger RNA translation or stability, nearly all small-RNA silencing pathways also seem to act at the DNA and chromatin level (Table 1). Studies in *Schizosaccharomyces pombe* (fission yeast) and other organisms suggest that small RNAs access DNA through interactions with nascent RNA transcripts, revealing a close relationship between nuclear and cytoplasmic RNA silencing mechanisms. Moreover, small-RNA silencing pathways seem to be intimately integrated with the RNA surveillance and processing pathways that determine the ultimate fate of RNA transcripts. Together, these studies reveal a broad and previously unsuspected role for RNAi and other RNA-processing mechanisms in the regulation of the structure and expression of eukaryotic genomes. Here, I discuss small-RNA silencing pathways and their role in chromatin regulation, drawing parallels between well-established examples in *S. pombe* and other organisms.

RNA silencing pathways

RNA silencing pathways can be broadly classified into different branches based on their mechanism of action, subcellular location and the origin of the small RNA molecules that they use (Table 1). However, the different branches have common components and intersect in some instances. siRNAs act in both the nucleus and the cytoplasm and are involved in PTGS and chromatin-dependent gene silencing (CDGS). CDGS refers to both transcriptional gene silencing (TGS) and co-transcriptional gene silencing (CTGS)[3]. miRNAs are generated from hairpin precursors by the successive actions of the RNaseIII enzymes Drosha and Dicer, which are located in the nucleus and cytoplasm, respectively (see page 396 for a more detailed discussion of small RNA precursor processing and complex assembly). Although Drosha is absent in plants, the general features of the miRNA pathway are conserved in plants and animals, but not in fungi and other protozoa. Whereas the vast majority of miRNAs seem to act exclusively in the cytoplasm and mediate mRNA degradation or translational arrest[19], some plant miRNAs may act directly in promoting DNA methylation[20]. Furthermore, recent studies describe a role for promoter-directed

[1]Howard Hughes Medical Institute, and Department of Cell Biology, Harvard Medical School, Boston, Massachusetts 02115, USA.

Table 1 | Conservation of small-RNA silencing pathways in eukaryotes

Small RNA	Size (nucleotides)	Mechanism of action	Eukaryotes conserved in
siRNA	~21–25	PTGS (RNA degradation or translational arrest) CDGS	Plants, animals, fungi, ciliates
miRNA	~21–25	PTGS (RNA degradation or translational arrest) CDGS (to a lesser extent)	Plants, animals
piRNA	~24–31*	PTGS (RNA degradation) CDGS (to a lesser extent)	Animals

All three of the major RNA silencing pathways identified thus far seem to act in both post-transcriptional gene silencing (PTGS) and chromatin-dependent gene silencing (CDGS) pathways. CDGS refers to chromatin-dependent silencing events that involve the assembly of small RNA complexes on nascent transcripts and includes both transcriptional gene silencing (TGS) and co-transcriptional gene silencing (CTGS) events. The latter involves the chromatin-dependent processing or degradation of the nascent transcript. *_Caenorhabditis elegans_ piRNAs are 21 nucleotides.

human miRNAs in facilitating repressive chromatin modifications and TGS[21,22]. siRNAs are generated from long dsRNA precursors, which can be produced from a variety of single-stranded RNA (ssRNA) precursors. These precursors include sense and antisense RNAs transcribed from convergent promoters, which can anneal to form dsRNA, and hairpin RNAs that result from transcription through inverted repeat regions[23–25] (Fig. 1a). In some situations the long dsRNA is produced enzymatically from certain aberrant or non-coding RNA precursors. One example of this pathway involves aberrant RNAs that lack processing signals or are produced by Argonaute slicer activity. These RNAs recruit RNA-dependent RNA polymerase (RdRP) enzymes, which recognize free 3′ ends and synthesize dsRNA[2,26,27] (Fig. 1b, c). Here RdRP enzymes are in competition with the TRAMP polyadenylation pathway, which targets aberrant RNAs for degradation by a 3′→5′ exonuclease complex, called the exosome[28–31](Fig. 1b). The siRNA branch of the pathway seems to be conserved from fungi to mammals (Table 1), although *Drosophila melanogaster* (fruitflies) and mammals lack RdRPs and cannot amplify siRNAs.

piRNAs originate from a diversity of sequences, including repetitive DNA and transposons, and like siRNAs they seem to act at both the post-transcriptional and chromatin levels[12–18]. The mechanism(s) that generates and amplifies piRNAs is not yet fully elucidated but involves the slicer activity of the PIWI-clade proteins themselves[4] (Fig. 1d). This class of small RNAs is present in *D. melanogaster*, *C. elegans* and mammals, but seems to be absent in fungi and plants (Table 1).

Small RNAs in DNA and chromatin regulation

An accumulating body of evidence supports an important role for small RNAs in the modulation of chromatin structure and TGS in plants, fungi and animal cells. RNA silencing was first linked to TGS by the discovery that transgene and viral RNAs guide the methylation of homologous DNA sequences in plants[32]. Analysis of the guide RNAs in *Arabidopsis thaliana* revealed that these RNAs were processed into small RNAs of ~25 nucleotides, similar to the size previously described for miRNAs[5,33]. This observation and the realization that exogenously introduced dsRNA in animals is processed into siRNAs[8] established small RNAs as central players in diverse RNA silencing pathways. Later studies in *A. thaliana* indicated that RNA-directed DNA methylation of the *FWA* transgene requires Dicer (DCL3) and Argonaute (AGO4), and is linked to histone H3 lysine 9 (H3K9) methylation, indicating that RNA-directed DNA methylation and RNAi have common molecular mediators[34–36].

Evidence for the role of RNA silencing in mediating changes at the chromatin level also came from studies of silent or heterochromatic DNA domains in unicellular eukaryotes, such as *S. pombe* and the ciliate *Tetrahymena thermophila*. *S. pombe* contains single genes encoding the Argonaute, Dicer and RdRP proteins, called *ago1*, *dcr1* and *rdp1*, respectively. Deletion of any of these genes results in loss of heterochromatic gene silencing, markedly reduced H3K9 methylation at centromeric repeats, and accumulation of non-coding RNAs, which are transcribed from centromeric repeat regions and processed into siRNAs[37,38]. Moreover, RNAi is directly linked to a structural component of heterochromatin through RITS, which in *S. pombe* contains Ago1, the chromodomain protein Chp1, the glycine and tryptophan (GW)-motif-containing protein Tas3 and centromeric siRNAs[10,29,39]. *T. thermophila* cells are binucleate with a germline micronucleus and a somatic macronucleus. Development of a new macronucleus after sexual conjugation and meiosis involves massive DNA elimination of non-genic sequences. This elimination requires TWI1, a *T. thermophila* PIWI-clade protein, and PDD1, a chromodomain protein that binds to both K9- and K27-methylated histone H3 (refs 40–42). In addition, DNA elimination is associated with Dicer-produced small RNAs, called scan RNAs (scnRNAs), giving rise to the idea that a scnRNA RITS-like complex targets sequences destined for elimination into heterochromatin[40]. However, a physical association between chromatin proteins and TWI1 has not yet been reported.

RNAi is also linked to chromatin modifiers in animal cells. In *D. melanogaster*, the introduction of multiple tandem copies of a transgene results in silencing of both the transgene array and the endogenous copies. This repeat-induced gene silencing, which is analogous to RNA-mediated co-suppression in plants[2], requires components of the Polycomb group (PcG) of genes, as well as several RNAi factors, including PIWI and AGO2 (ref. 43). The PcG gene products are chromatin-binding and -modifying repressors that prevent the expression of homeobox (HOX) regulators outside their proper domains of expression[44]. The requirement for both PcG proteins and PIWI in transgene silencing suggested the possibility that in *D. melanogaster*, as in plant cells, RNA silencing could operate at the chromatin level. In fact, later studies showed that RNA silencing factors are also required for the formation of *D. melanogaster* centric heterochromatin, recruitment of heterochromatin protein 1 (HP1) and silencing of transgenes that are inserted in pericentromeric heterochromatin[43,45]. In addition to HP1 and PIWI, efficient silencing requires DCR-1, PIWI, Aubergine (AUB) and the putative helicase HLS (also known as SPN-E)[43]. Moreover, silencing of a mini-*white* gene, which is mediated by a *cis*-acting repeated element from the heterochromatic Y chromosome, requires HP1, SU(VAR)3-9 (the H3K9 methyltransferase), as well as PIWI, AUB, HLS and DCR-1 (ref. 46). Transgene-induced gene silencing in *C. elegans* has also been shown to require RNAi and chromatin modifiers[47,48]. Surprisingly, screens for defects in classical RNAi, mediated by feeding of dsRNA, have also uncovered several chromatin modifiers, suggesting that perhaps the connection between RNAi and chromatin modifiers may not be limited to repeat-induced silencing[49].

In contrast to their apparent requirement for PcG-mediated repeat-induced gene silencing, RNAi components do not seem to be required for PcG-mediated silencing of HOX genes outside their proper domains of expression[50]. Mutations in several RNA silencing factors disrupt the silencing of a tandem mini-*white* gene array and perturb the nuclear clustering of PcG-repressed HOX loci[50]. However, despite their requirement for PcG-mediated repeat-induced silencing, loss of PIWI and RNAi components does not lead to a loss of HOX gene silencing. The simplest explanation for these observations is that RNAi is required for some, but not all, PcG-mediated silencing events.

Linking heterochromatin to RNAi

Heterochromatin is associated with repetitive DNA sequences and transposons, and has important roles in chromosome transmission, maintenance of genomic stability, and regulation of gene expression[51–53]. With the exception of budding yeast, which lacks centromeric DNA repeats, heterochromatin is concentrated at repeats and transposons that surround centromeres, telomeres and other genomic loci (Fig. 2a). Two important defining properties of heterochromatin involve its modes of assembly and inheritance. First, heterochromatin assembly involves nucleation sites, which act as entry points for the recruitment and spreading of repressor proteins. Unlike recruitment, which involves the action of a site-specific DNA-binding protein or RNA molecule,

spreading occurs in a sequence-independent manner and involves changes in chromatin structure that are mediated by histone-modifying enzymes. The second defining property of heterochromatin is its mode of inheritance. Once assembled, heterochromatin is inherited through many cell divisions, at least partly independently of the underlying DNA sequence. The mechanisms of spreading and epigenetic inheritance of heterochromatin are poorly understood, but, in *S. pombe*, require components of the RNAi pathway[3,53,54].

At the molecular level, heterochromatin is characterized by association with hypoacetylated histones and, in organisms ranging from *S. pombe* to humans, by association with H3K9 dimethylation and trimethylation[3,51]. H3K9 is methylated by SU(VAR)3-9 in *D. melanogaster*, SUV39H in humans and Clr4 in *S. pombe*, and creates a binding site for HP1 (Swi6 and Chp2 in *S. pombe*)[55–58]. HP1 proteins contain a chromodomain that binds to methylated H3K9 and a chromoshadow (CSD) domain, which is involved in other protein–protein interactions[54].

Biochemical isolation of *S. pombe* heterochromatin and RNAi complexes has provided direct physical links between heterochromatin and RNAi proteins, leading to models for how RNAi mediates heterochromatin assembly and participates in gene silencing. In addition to HP1 proteins, heterochromatic gene silencing in *S. pombe* requires the chromodomain protein Chp1 (ref. 59). Chp1 is larger than HP1 and, like the Polycomb (Pc) subfamily of chromodomains, contains only a single chromodomain at its amino terminus. Like Swi6 and Chp2, Chp1 is a structural component of heterochromatin and is required for heterochromatic gene silencing[59]. Unlike Swi6 and Chp2, which are not required for H3K9 methylation within centromeric repeat regions[58,60], a lack of Chp1 in cells leads to a marked loss of H3K9 methylation, indicating that Chp1 has a critical role in heterochromatin formation.

Biochemical purification of Chp1 showed that it is associated with Ago1 in RITS[10]. RITS acts as a specificity determinant for the recruitment of other RNAi complexes and chromatin-modifying enzymes to specific DNA regions. RITS also physically associates with and is required for recruitment of the RNA-directed RNA polymerase complex (RDRC) to non-coding RNAs that are transcribed from centromeric repeats[61,62]. RDRC contains the *S. pombe* RNA-directed RNA polymerase, Rdp1, a putative helicase termed Hrr1, and Cid12, a member of the Trf4 and Trf5 family of polyadenylation polymerases[61], which were first identified in the budding yeast *Saccharomyces cerevisiae* and are involved in the degradation of aberrant transcripts[30,31]. The physical association of RITS and RDRC is siRNA- and Clr4-dependent, suggesting that this association occurs on chromatin and requires histone H3K9 methylation[61]. These observations further suggest that RITS and RDRC localize to chromatin-bound nascent RNA by a mechanism that involves tethering the nascent transcript to chromatin via the bivalent complex RITS (Fig. 2). In addition to RITS, *S. pombe* contains a second Ago1-containing complex, named Argonaute siRNA chaperone (ARC) complex[63]. The Ago1 protein in the ARC complex contains duplex, rather

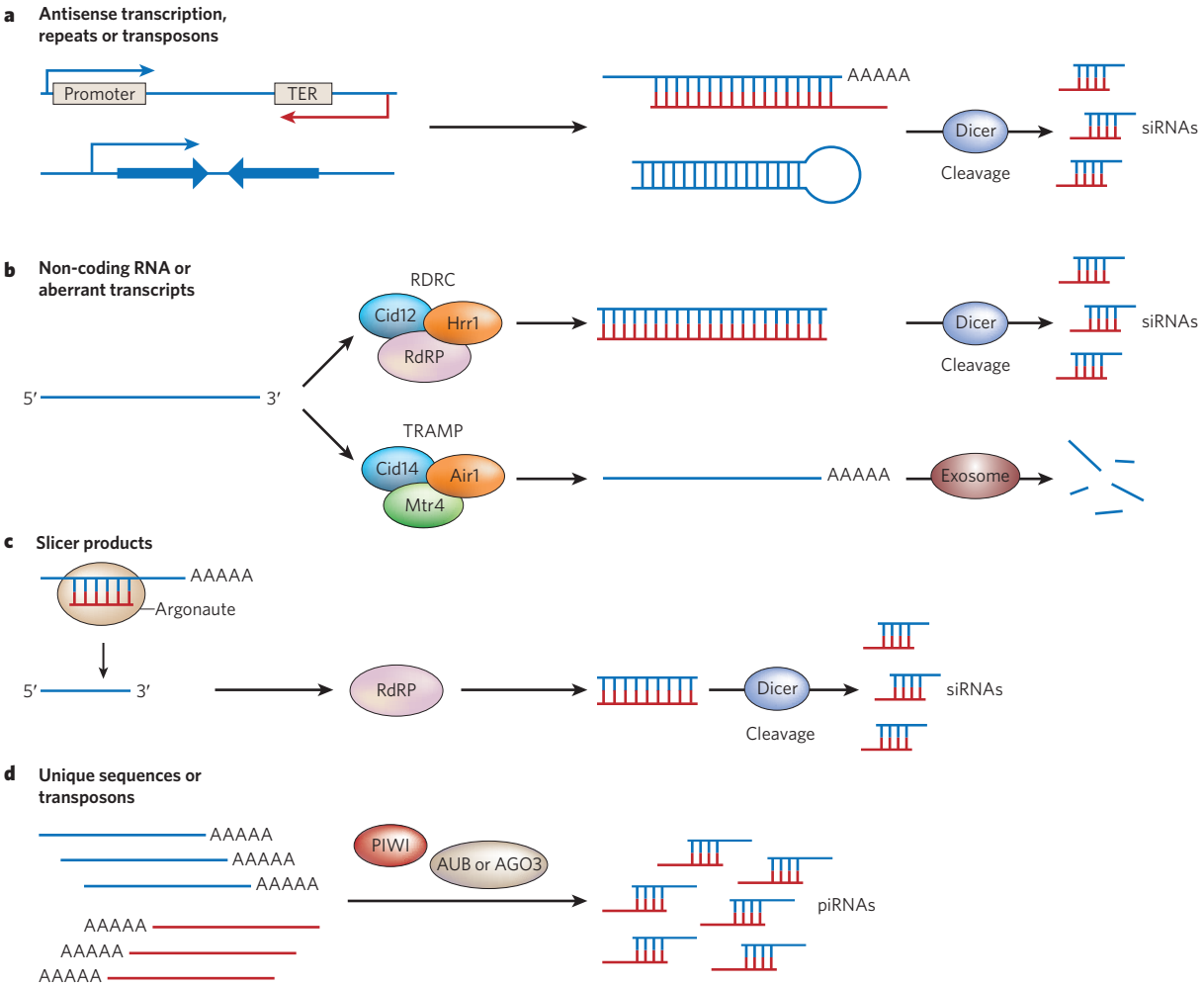


Figure 1 | Pathways of RNA processing and biogenesis of small RNAs. **a**, Generation of endogenous siRNAs from dsRNA resulting from convergent transcription (sense–antisense RNA base-pairing; top) or transcription through inverted repeat sequences (hairpin RNA formation; bottom). TER, transcription termination signal. **b**, Processing of non-coding and aberrant RNAs by the RDRC and TRAMP complexes, containing the Cid12 and Cid14 non-canonical polyadenylation polymerases, respectively; the RDRC/Dicer pathway produces duplex siRNAs, whereas the TRAMP/exosome pathway produces single-stranded degradation products. **c**, Generation of a free 3′ end by the slicer activity of an Argonaute protein, which can be processed into dsRNA by RdRP or targeted for degradation by the exosome (not shown). **d**, Pathway for the generation of piRNAs by the PIWI clade of Argonaute proteins: PIWI, AUB and AGO3.

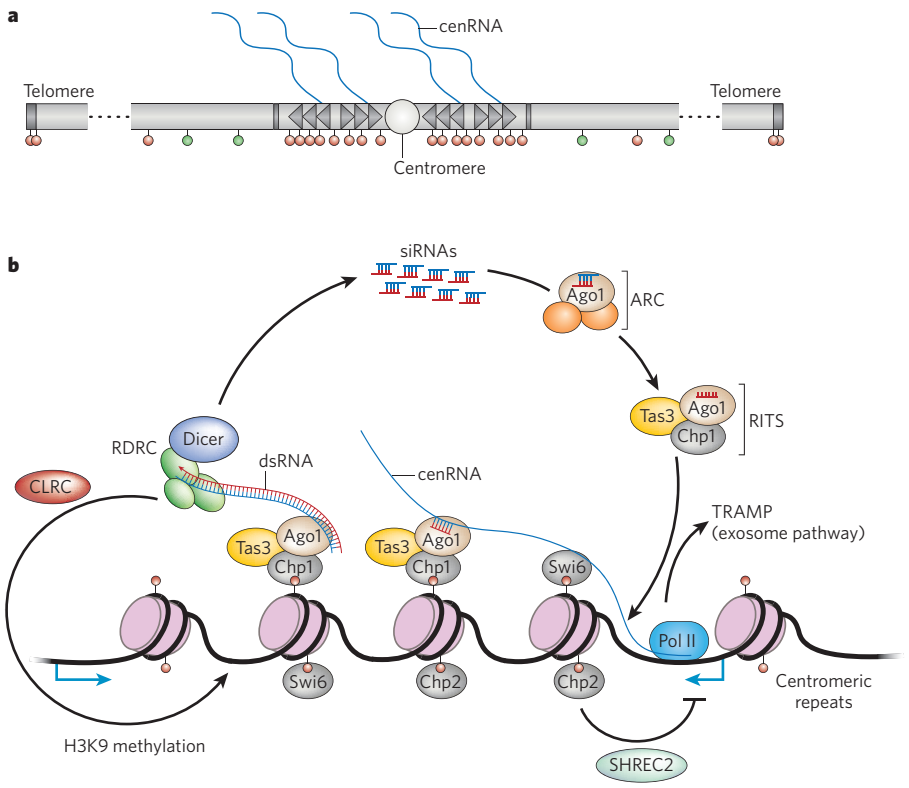


Figure 2 | Chromosome organization and the nascent transcript model for heterochromatic gene-silencing assembly in *Schizosaccharomyces pombe*. a, The structure of *S. pombe* centromeric repeat regions, highlighting the presence of non-coding centromeric transcripts (cenRNA) and association with histone H3 that is dimethylated and trimethylated on lysine 9 (red lollipops) as opposed to histone H3 that is methylated on lysine 4 (green lollipops) in euchromatic regions. **b**, The nascent transcript model for heterochromatin assembly. The RITS is tethered to chromatin through base-paring interactions between siRNAs and nascent non-coding transcripts and interactions with H3K9-methylated nucleosomes, resulting in the recruitment of RDRC–Dicer, dsRNA synthesis and siRNA amplification. This RNAi positive-feedback loop then recruits the CLRC H3K9 methyltransferase. Efficient silencing also requires two HP1 proteins (Swi6 and Chp2), which promote the association of RITS with the non-coding RNA or mediate TGS through recruitment of the SHREC2 deacetylase complex, respectively. Another tier of regulation, involving the degradation of heterochromatic transcripts by the TRAMP/exosome pathway, further ensures full gene silencing. Blue arrows (bottom) highlight convergent transcription resulting in synthesis of sense and antisense RNAs, which may contribute to the production of trigger siRNAs.

than single-stranded, siRNA, indicating that the slicer activity of Ago1 (refs 63, 64), which is required for the release of the siRNA passenger strand, is inhibited in this complex[63].

Nascent transcripts as assembly platforms

In principle, siRNAs in RITS can base-pair with either unwound DNA regions or with nascent non-coding RNAs that are transcribed from their target DNA. The two models are not mutually exclusive and base-pairing with DNA and RNA may contribute to different aspects of the mechanism of siRNA biogenesis and function. However, although a role for siRNA–DNA base-pairing cannot be ruled out at this point, several lines of evidence support siRNA–RNA base-pairing interactions in which the siRNA targets nascent non-coding transcripts (Fig. 2). First, RITS associates with the RDRC, which uses ssRNA as a template to synthesize dsRNA, providing evidence that RITS itself is RNA-associated[61]. Furthermore, the RITS–RDRC interaction requires siRNA and the Clr4 H3K9 methyltransferase, suggesting that it occurs on heterochromatin-bound transcripts[61]. Together with the observation that proteins required for heterochromatin formation — such as Sir2, Swi6, Clr4 and other components of the Clr4 methyltransferase complex (CLRC), as well as RITS and RDRC — are required for siRNA accumulation[10,61,65–67], these studies suggest that siRNA-programmed RITS localizes to nascent chromatin-tethered non-coding transcripts and recruits the RDRC to initiate dsRNA synthesis and siRNA amplification (Fig. 2b). Direct support for this model comes from experiments in which the Tas3 component of RITS was fused to the phage λ N (λN) protein and tethered to the transcript of a euchromatic *ura4*$^+$ gene, which was modified with the addition of five λN-binding sites upstream of its transcription termination sequences (*ura4-5BoxB*)[66]. In cells containing *ura4-5BoxB*, the Tas3-λN protein could efficiently initiate *de novo* siRNA generation and heterochromatin formation[66]. Like the situation at centromeres, siRNA generation in this system is H3K9 methylation-dependent, suggesting that Tas3-λN associates with chromatin-bound nascent transcripts and initiates RNAi-mediated heterochromatin assembly. Finally, several splicing factors associate with RDRC[61,68] and are required for RNAi-mediated centromeric gene silencing[68]. These results provide additional support for co-transcriptional processing of non-coding centromeric RNAs, as spliceosomal components are known to associate with nascent RNA transcripts co-transcriptionally. A role for the nascent transcript in acting as a template for the recruitment of chromatin-modifying activities may be conserved throughout eukaryotes. For example, large non-coding RNAs such as XIST, which is involved in X-chromosome inactivation, are thought to be involved in the recruitment of histone and DNA methyltransferase enzymes[69]. However, the mechanism of recruitment of chromatin-modifying activities to XIST may involve site-specific RNA-binding proteins rather than small RNAs.

Chromatin-dependent processing of siRNAs

A remarkable observation in studies of RNAi in *S. pombe* is that the generation of centromeric siRNAs is a heterochromatin-dependent event[61,65]. In the nascent transcript model (Fig. 2b), RITS associates with methylated H3K9 through the chromodomain of its Chp1 component and captures the nascent non-coding transcript through base-pairing interactions involving siRNAs bound to its Ago1 protein. In cells lacking the H3K9 methyltransferase Clr4 or any component of the CLRC, the levels of centromeric siRNAs are greatly diminished[61,67]. Moreover, one of the two HP1 proteins, Swi6, is required for efficient siRNA generation[61,66,70] and the association of RDRC with centromeric DNA repeats[62] and non-coding centromeric RNAs[71]. Furthermore, the crucial chromatin-dependent step in siRNA generation involves dsRNA synthesis by RDRC, as the introduction of a long dsRNA-containing hairpin into *S. pombe* cells circumvents the requirement for both RDRC and Clr4 in siRNA generation[70]. These results suggest that RDRC is only able to synthesize dsRNA on chromatin-bound templates after it has been recruited by RITS, revealing the existence of a chromatin-dependent step in the activation of the dsRNA biogenesis and siRNA amplification pathway in *S. pombe*. The resultant dsRNA is processed into siRNA by Dcr1, which is also physically tethered to RDRC[72]. Heterochromatin regulation of small RNA production may be conserved in metazoans. X-TAS (transposable *P* elements inserted in telomeric-associated sequences on the X chromosome) and *flamenco*, two major piRNA-producing loci that control the transposition of *P* and *gypsy* elements in *D. melanogaster*, respectively, are embedded

in heterochromatin, and their genome defence function requires both PIWI and HP1 (refs 73, 74).

siRNA-mediated initiation of chromatin silencing

An important question regarding the role of RNA in gene silencing is whether small RNAs can initiate *de novo* chromatin modifications. Although small RNAs are important components of some CDGS mechanisms, their ability to initiate chromatin modifications seems to be under strict control by other mechanisms. In *S. pombe*, ectopically produced hairpin siRNAs can initiate H3K9 methylation and gene silencing at only a subset of target loci[70]. siRNA-mediated CDGS correlates with chromosomal location and the occurrence of antisense transcription at the targeted locus, and requires overexpression of the Swi6 (HP1) protein[70]. This may be reflecting the importance of cooperativity in the recruitment of RITS and other Argonaute or PIWI effector complexes to chromatin. In addition to siRNAs, stable association of RITS with chromatin requires the binding of the chromodomain in Chp1 to H3K9-methylated nucleosomes[10,65]. In the absence of H3K9 methylation, the initial binding of RITS to chromatin may be inefficient. Swi6 overproduction may help in initial RITS binding by stabilizing low levels of H3K9 methylation that occur throughout the genome, or alternatively by tethering the nascent transcript at the target locus to chromatin[71] (Fig. 2b). Similar limitations may explain the context-dependent ability of siRNAs to promote DNA methylation in plants[75], as well as the observed variability in siRNA-mediated chromatin modifications in animal cells (for example, see refs 76, 77). The ability of siRNAs to act as initiators is reminiscent of the role of DNA-binding transcription factors in the regulation of transcription, which often involves cooperativity between two or more transcription factors and is sensitive to local chromatin structure.

Small RNAs and epigenetic inheritance

Mechanisms that mediate the *cis*-inheritance of chromatin states and their associated gene-expression patterns remain enigmatic. It has long been known that during DNA replication, old parental histones are randomly distributed onto the two newly synthesized daughter DNA strands[78]. This retention of old histones during DNA replication has given rise to the idea that histone modifications mediate the epigenetic inheritance of chromatin states. Histone modifications, such as H3K9 methylation, create binding sites for proteins such as Chp1, Chp2 and Swi6, as well as the methyltransferase Clr4 (ref. 79) (Fig. 2b). Their retention during DNA replication could in principle serve as a mark for the re-recruitment of new chromatin-modifying activities that re-establish old modification patterns. However, the affinity of modified histones for specific binding proteins may be too low to allow the specific re-establishment of chromatin states, and other inputs into the mechanism are required[44]. The nascent transcript model, described above, provides a possible mechanism for epigenetic inheritance of heterochromatin. As in plants and other systems that contain an RdRP-dependent siRNA amplification mechanism[2,80], the siRNA generation mechanism in *S. pombe* is likely to form a positive-feedback loop[61,62]. Two specific features of this loop may underlie the mechanism that ensures the epigenetic inheritance of histone H3K9 methylation and heterochromatin. First, siRNAs can recruit H3K9 methylation to chromatin, possibly through physical interactions with the CLRC or dsRNA[70,81]. Thus, so long as siRNAs corresponding to a specific chromatin domain are present, they can recruit H3K9 methylation to that domain (Fig. 2b). The second feature involves a requirement for H3K9 methylation and chromatin localization in activating the siRNA positive-feedback loop[61,62,65]. This ensures that siRNAs are *cis*-restricted and is central to the role of siRNAs as epigenetic maintenance factors: siRNAs act only on those daughter DNA strands that have inherited old parental histone H3 molecules containing H3K9 methylation. Such cooperativity-based mechanisms involving the dual recognition of histone marks and other specificity factors (siRNAs or DNA-binding proteins) are likely to underlie all epigenetic *cis*-inheritance mechanisms.

RNAi and exosome-mediated RNA degradation

It may seem paradoxical that RNAi, which requires transcription, is required for assembling heterochromatin, a state that is associated with gene inactivation and TGS[3,51]. However, multiple mechanisms seem to ensure that transcription in heterochromatin does not result in the production of mature transcripts, thereby keeping heterochromatic genes off, despite transcription. First, heterochromatic transcripts are degraded or processed into siRNAs by the RNAi machinery itself through a process that has been referred to as CTGS or *cis*-PTGS[65,66] (Fig. 2b). CTGS requires the tethering of the RNAi machinery to heterochromatin by H3K9 methylation. This mechanism makes a major contribution to the silencing of some promoters in centromeric DNA repeats, although TGS is also an important contributing mechanism[66,71,82]. Second, an RNAi-independent RNA surveillance mechanism involving the TRAMP polyadenylation complex, which contains Cid14 (a Trf4/5 homologue), Air1, and Mtr1 in *S. pombe*, also targets heterochromatic transcripts for degradation[28]. In *S. cerevisiae*, TRAMP recognizes aberrant transcripts that lack polyadenylation signals and targets them for degradation by the exosome, a 3′→5′ exonuclease complex[30,31,83]. The presence of another member of the Trf4 polyadenylation polymerase family, Cid12, in the RDRC[61] suggests that RDRC and TRAMP may compete for access to heterochromatic transcripts (Fig. 1b). Furthermore, TRAMP and RDRC may compete more broadly for RNA substrates, because in *cid14* deletion cells new classes of RNAs become RNAi targets and are processed into siRNAs[29]. The involvement of members of the Trf4 family in RNAi processes in *C. elegans* and *T. thermophila* suggests a conserved role for members of this family in the coordination of exosome-mediated RNA surveillance with RNAi[84,85]. Finally, transcription in heterochromatin is cell-cycle regulated and is largely restricted to the S phase of the cell cycle[82,86]. This transcription is associated with high levels of siRNAs during the S phase, which may be important for epigenetic re-establishment of histone H3K9 methylation by the RITS–RDRC–CLRC complexes. However, it remains to be determined whether the increase in centromeric transcription and siRNA levels in S phase is merely a reflection of cell-cycle-associated changes in chromatin structure or has an important role in RNAi-mediated heterochromatin assembly.

Nearly all co-transcriptional RNA-processing events studied so far, including pre-mRNA capping, splicing and 3′-end processing, involve association between components of the processing machinery and RNA polymerase II (Pol II). Association with the polymerase is thought to help ensure that processing occurs in an orderly fashion and couples mRNA maturation with mRNA export. In addition, these associations serve to couple RNA quality control with transcription, ensuring that only true mRNAs are exported from the nucleus for translation. There is evidence that RNAi-mediated co-transcriptional heterochromatin assembly also involves interactions with Pol II[87,88]. Point mutations in two different Pol II subunits in *S. pombe*, Rpb2 and Rpb7, have been isolated in screens for defects in centromeric heterochromatin assembly. Neither mutation is associated with a growth defect or general perturbation of transcription[87,88], suggesting that the mutations may affect specific interactions with components of the RNAi machinery or the CLRC. Such interactions may contribute to efficient siRNA generation or H3K9 methylation by stabilizing the association of RITS–RDRC with nascent transcripts. Interestingly, in *A. thaliana*, RNA-dependent DNA methylation involves interactions between an Argonaute protein and RNA Pol IV, a plant-specific DNA-dependent RNA polymerase[89] (discussed below).

Conservation of small-RNA-mediated silencing

As discussed above, RNA silencing mechanisms have critical roles in endogenous chromatin-mediated processes in plants, *C. elegans*, *D. melanogaster*, ciliates and fungi. The role of small RNAs in chromatin silencing can also be extended to mammalian cells, although the mechanisms and physiological pathways are not yet clear. Reports from several laboratories provide evidence for the occurrence of DNA and histone modifications, which are promoted by the introduction

of siRNAs or hairpin RNAs into mammalian cell lines[76,77,90,91]. In these studies, siRNAs are directed to the promoter regions of target genes and induce the recruitment of repressive histone marks such as H3K9 and H3K27 methylation[91], but silencing is not always associated with CpG methylation[76]. In addition to chromatin-modifying complexes, siRNA-mediated TGS in mammalian cells requires AGO1 and AGO2 when the gene encoding progesterone is targeted[77], and AGO1 when the gene encoding the human immunodeficiency virus 1 co-receptor (CCR5) is targeted[91]. The recent identification of endogenous siRNAs in *D. melanogaster* and mammalian cells, which map to intergenic regions and are produced from dsRNA resulting from antisense transcription or long-hairpin structures, raises the intriguing possibility that some of these siRNAs modulate chromatin structure[23–25].

The PIWI-clade proteins and their associated piRNAs have important roles in the control of transposons in the germline — and possibly somatic cells — of *D. melanogaster* and mammals[4]. The mouse MIWI2 member of this family is required for silencing the long interspersed nuclear element 1 (*LINE-1*) and intracisternal A particle (*IAP*) transposable elements in the testis, and in *Miwi2* mutants both *LINE-1* and *IAP* DNA is demethylated[92], suggesting that piRNAs, directly or indirectly, mediates changes in DNA methylation. It remains unclear how the role of PIWI proteins in transposon silencing in the germline may be related to their function in repeat-induced and heterochromatic gene silencing in somatic cells described in *D. melanogaster*[45,46].

Although the mechanisms that link RNA to chromatin and the biochemical nature of the relevant complexes have not been defined yet, the available evidence allows us to draw some parallels between the nascent transcript model in *S. pombe* and other systems. The common denominator in the RNA silencing pathways operating in genome regulation is the linkage of Argonaute or PIWI proteins to chromatin- or DNA-associated molecules (Fig. 3). Argonaute proteins associate with adaptor proteins containing the conserved GW motif, which binds to their PIWI domain and is required for miRNA-mediated silencing[93] (Fig. 3a). In *S. pombe*, the GW-motif-containing protein Tas3 links Ago1 to Chp1 (refs 10, 94, 95). The binding of Chp1 to a methylated nucleosome then serves to tether nascent non-coding RNA, which is base-paired with siRNA in Ago1, to chromatin (Fig. 3b). This tethering seems to be crucial in that it links RNAi to chromatin and 'activates' the Ago1-bound nascent transcript complex to mediate chromatin modifications[61,65]. A similar Argonaute tethering situation seems to exist in *A. thaliana*, where, in addition to AGO4 and DCL3, RNA-directed DNA methylation requires the plant-specific Pol IV[96–99]. Pol IV exists as Pol IVA and Pol IVB complexes, which differ in their largest component, NRPD1A and NRPD1B, respectively. Pol IVB and AGO4 are thought to act downstream of Pol IVA and DCL3, which are required for siRNA generation, to trigger DNA methylation. NRPD1B contains a GW motif at its carboxyl terminus[89]. This GW-motif-containing domain links Pol IVB to AGO4, providing a parallel with the function of other GW-domain-containing proteins, such as the Tas3 component of RITS in *S. pombe*[89,96] (Fig. 3e). Thus, in plants, the strategy for coupling RNAi to chromatin involves a physical interaction between a repeat- or heterochromatin-specific RNA polymerase and an Argonaute protein. Once an siRNA-programmed AGO4 localizes to a nascent transcript synthesized by Pol IVB, it may trigger histone H3K9 and DNA methylation by recruiting the appropriate methyltransferase enzymes (Fig. 3e).

The role of the *D. melanogaster* PIWI protein in repeat-induced gene silencing and heterochromatin assembly seems to involve a direct association between PIWI and HP1 (ref. 100) (Fig. 3c). PIWI–HP1 may function as a RITS that targets nascent transcripts in repeat DNA elements and tethers these transcripts to chromatin by means of base-pairing interaction between piRNAs in PIWI and the association of PIWI with HP1 (Fig. 3c). Unlike the case with RITS and AGO4, this tethering does not seem to involve a GW-domain-containing protein and is mediated by the HP1 CSD and a conserved CSD-binding PXVXL motif (where X

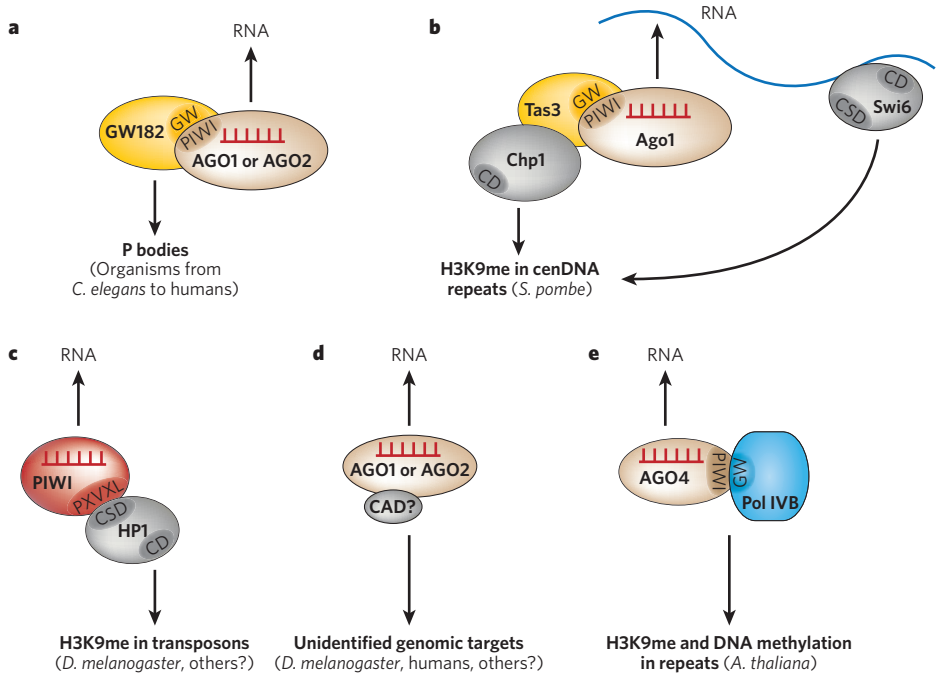


Figure 3 | Argonaute complexes that link RNA silencing to chromatin modifications. Argonaute proteins in different silencing pathways, including miRNA- and siRNA-mediated PTGS, are associated with conserved GW-motif-containing adaptor proteins, which help direct them to different targets. **a**, In many organisms, GW182 (a GW-motif-containing protein) or one of its homologues associates with the AGO1 and AGO2 proteins and directs them to P bodies. **b**, In *S. pombe*, Ago1 in the RITS is linked to heterochromatin through its association with the GW protein Tas3, which also binds to Chp1. Chp1 in turn associates with H3K9 methylated nucleosomes (H3K9me) through its chromodomain (CD). Swi6 (a homologue of HP1) acts as an accessory factor that helps tether the non-coding RNA to heterochromatin. The chromoshadow domain (CSD) is involved in protein–protein interactions. cenDNA, centromeric repeat DNA. **c**, In *D. melanogaster*, PIWI is targeted to heterochromatin through direct interactions with HP1; the association of PIWI with HP1 is mediated through the PXVXL motif, present in many HP1-binding proteins, rather than through a GW motif. **d**, In *D. melanogaster* and possibly other organisms, AGO1 and AGO2 have been implicated in mediating chromatin modifications, but the putative chromatin adaptor (CAD) protein has not been identified. **e**, In *A. thaliana*, AGO4 is linked to Pol IVB, which contains a GW motif at its carboxyl terminus and is specifically required for DNA methylation and silencing of heterochromatic repeats.

is any amino acid) present in *D. melanogaster* PIWI[100]. The PIWI–HP1 complex may be required for the recruitment and spreading of H3K9 methylation or possibly for the co-transcriptional degradation of RNAs that may escape heterochromatic TGS. It remains to be determined whether this PIWI–HP1 complex acts more broadly in piRNA-mediated silencing of transposons in the germline. Similarly, the possible role of AGO1 and AGO2 in siRNA-dependent gene silencing may be mediated by interactions with unidentified chromatin adaptors (Fig. 3d).

Future prospects

RNAi and related RNA silencing pathways have emerged as new mechanisms for the regulation of the structure and activity of genes and genomes. Our understanding of the mechanisms that allow some small RNAs to act at the DNA and chromatin level, and restrict other small RNAs to mRNA regulation in the cytoplasm, is still at an early stage. Although accumulating evidence suggests that nuclear small-RNA pathways are conserved, the endogenous pathways that may use small RNAs for genome regulation in animal cells remain for the most part unknown. Another gap in our knowledge of nuclear small-RNA pathways in animal cells involves the biochemical identification of the molecular networks that link different types of small RNA to chromatin proteins. Whereas Argonaute and PIWI proteins, as well as small RNAs, have been implicated in mediating chromatin or DNA modifications, it remains unclear how specific chromosome regions are targeted and how modifying enzymes are recruited. Future studies are likely to provide new and surprising insights about the way in which small and large non-coding RNAs regulate chromatin structure and how this ability is, in turn, regulated. ■

1. Fire, A. *et al.* Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391,** 806–811 (1998).
2. Baulcombe, D. RNA silencing in plants. *Nature* **431,** 356–363 (2004).
3. Buhler, M. & Moazed, D. Transcription and RNAi in heterochromatic gene silencing. *Nature Struct. Mol. Biol.* **14,** 1041–1048 (2007).
4. Aravin, A. A., Hannon, G. J. & Brennecke, J. The Piwi–piRNA pathway provides an adaptive defense in the transposon arms race. *Science* **318,** 761–764 (2007).
5. Hamilton, A. J. & Baulcombe, D. C. A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science* **286,** 950–952 (1999).
6. Zamore, P. D., Tuschl, T., Sharp, P. A. & Bartel, D. P. RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell* **101,** 25–33 (2000).
7. Bernstein, E., Caudy, A. A., Hammond, S. M. & Hannon, G. J. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* **409,** 363–366 (2001).
8. Elbashir, S. M., Lendeckel, W. & Tuschl, T. RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev.* **15,** 188–200 (2001).
9. Hammond, S. M., Bernstein, E., Beach, D. & Hannon, G. J. An RNA-directed nuclease mediates post-transcriptional gene silencing in *Drosophila* cells. *Nature* **404,** 293–296 (2000).
   This study describes the biochemical isolation of RISC.
10. Verdel, A. *et al.* RNAi-mediated targeting of heterochromatin by the RITS complex. *Science* **303,** 672–676 (2004).
    This study describes the purification and component identification of RITS, which physically links RNAi to heterochromatin.
11. Carmell, M. A., Xuan, Z., Zhang, M. Q. & Hannon, G. J. The Argonaute family: tentacles that reach into RNAi, developmental control, stem cell maintenance, and tumorigenesis. *Genes Dev.* **16,** 2733–2742 (2002).
12. Aravin, A. *et al.* A novel class of small RNAs bind to MILI protein in mouse testes. *Nature* **442,** 203–207 (2006).
13. Girard, A., Sachidanandam, R., Hannon, G. J. & Carmell, M. A. A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature* **442,** 199–202 (2006).
14. Grivna, S. T., Beyret, E., Wang, Z. & Lin, H. A novel class of small RNAs in mouse spermatogenic cells. *Genes Dev.* **20,** 1709–1714 (2006).
15. Lau, N. C. *et al.* Characterization of the piRNA complex from rat testes. *Science* **313,** 363–367 (2006).
16. Watanabe, T. *et al.* Identification and characterization of two novel classes of small RNAs in the mouse germline: retrotransposon-derived siRNAs in oocytes and germline small RNAs in testes. *Genes Dev.* **20,** 1732–1743 (2006).
17. Batista, P. J. *et al.* PRG-1 and 21U-RNAs interact to form the piRNA complex required for fertility in *C. elegans*. *Mol. Cell* **31,** 67–78 (2008).
18. Das, P. P. *et al.* Piwi and piRNAs act upstream of an endogenous siRNA pathway to suppress Tc3 transposon mobility in the *Caenorhabditis elegans* germline. *Mol. Cell* **31,** 79–90 (2008).
19. Filipowicz, W., Jaskiewicz, L., Kolb, F. A. & Pillai, R. S. Post-transcriptional gene silencing by siRNAs and miRNAs. *Curr. Opin. Struct. Biol.* **15,** 331–341 (2005).
20. Bao, N., Lye, K. W. & Barton, M. K. MicroRNA binding sites in *Arabidopsis* class III HD-ZIP mRNAs are required for methylation of the template chromosome. *Dev. Cell* **7,** 653–662 (2004).
21. Gonzalez, S., Pisano, D. G. & Serrano, M. Mechanistic principles of chromatin remodeling guided by siRNAs and miRNAs. *Cell Cycle* **7,** 2601–2608 (2008).
22. Kim, D. H., Saetrom, P., Snove, O. Jr & Rossi, J. J. MicroRNA-directed transcriptional gene silencing in mammalian cells. *Proc. Natl Acad. Sci. USA* **105,** 16230–16235 (2008).
23. Ghildiyal, M. *et al.* Endogenous siRNAs derived from transposons and mRNAs in *Drosophila* somatic cells. *Science* **320,** 1077–1081 (2008).
24. Chung, W. J., Okamura, K., Martin, R. & Lai, E. C. Endogenous RNA interference provides a somatic defense against *Drosophila* transposons. *Curr. Biol.* **18,** 795–802 (2008).
25. Kawamura, Y. *et al. Drosophila* endogenous small RNAs bind to Argonaute 2 in somatic cells. *Nature* **453,** 793–797 (2008).
26. Dalmay, T., Hamilton, A., Rudd, S., Angell, S. & Baulcombe, D. C. An RNA-dependent RNA polymerase gene in *Arabidopsis* is required for posttranscriptional gene silencing mediated by a transgene but not by a virus. *Cell* **101,** 543–553 (2000).
27. Sijen, T. *et al.* On the role of RNA amplification in dsRNA-triggered gene silencing. *Cell* **107,** 465–476 (2001).
28. Buhler, M., Haas, W., Gygi, S. P. & Moazed, D. RNAi-dependent and -independent RNA turnover mechanisms contribute to heterochromatic gene silencing. *Cell* **129,** 707–721 (2007).
    This study identified a role for co-transcriptional RNA processing, mediated by the TRAMP/exosome pathway, as an additional layer of regulation that is required for efficient heterochromatic gene silencing.
29. Buhler, M., Spies, N., Bartel, D. P. & Moazed, D. TRAMP-mediated RNA surveillance prevents spurious entry of RNAs into the *Schizosaccharomyces pombe* siRNA pathway. *Nature Struct. Mol. Biol.* **15,** 1015–1023 (2008).
30. Wyers, F. *et al.* Cryptic Pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. *Cell* **121,** 725–737 (2005).
31. LaCava, J. *et al.* RNA degradation by the exosome is promoted by a nuclear polyadenylation complex. *Cell* **121,** 713–724 (2005).
32. Wassenegger, M., Heimes, S., Riedel, L. & Sanger, H. L. RNA-directed *de novo* methylation of genomic sequences in plants. *Cell* **76,** 567–576 (1994).
33. Mette, M. F., Aufsatz, W., van der Winden, J., Matzke, M. A. & Matzke, A. J. Transcriptional silencing and promoter methylation triggered by double-stranded RNA. *EMBO J.* **19,** 5194–5201 (2000).
    These two studies described a role for RNA in promoting the methylation of homologous genomic sequences in plants.
34. Zilberman, D., Cao, X. & Jacobsen, S. E. *ARGONAUTE4* control of locus-specific siRNA accumulation and DNA and histone methylation. *Science* **299,** 716–719 (2003).
35. Henderson, I. R. *et al.* Dissecting *Arabidopsis thaliana* DICER function in small RNA processing, gene silencing and DNA methylation patterning. *Nature Genet.* **38,** 721–725 (2006).
36. Lippman, Z., May, B., Yordan, C., Singer, T. & Martienssen, R. Distinct mechanisms determine transposon inheritance and methylation via small interfering RNA and histone modification. *PLoS Biol.* **1,** e67 (2003).
37. Volpe, T. A. *et al.* Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science* **297,** 1833–1837 (2002).
    This study identified a role for components of the RNAi pathway in heterochromatin assembly and gene silencing at *S. pombe* centromeres.
38. Reinhart, B. J. & Bartel, D. P. Small RNAs correspond to centromere heterochromatic repeats. *Science* **297,** 1831 (2002).
39. Cam, H. P. *et al.* Comprehensive analysis of heterochromatin- and RNAi-mediated epigenetic control of the fission yeast genome. *Nature Genet.* **37,** 809–819 (2005).
40. Mochizuki, K., Fine, N. A., Fujisawa, T. & Gorovsky, M. A. Analysis of a *piwi*-related gene implicates small RNAs in genome rearrangement in *Tetrahymena*. *Cell* **110,** 689–699 (2002).
    This study identified a role for *T. thermophila* TWI1, a PIWI-clade Argonaute protein, and small RNAs in mediating DNA elimination.
41. Taverna, S. D., Coyne, R. S. & Allis, C. D. Methylation of histone H3 at lysine 9 targets programmed DNA elimination in *Tetrahymena*. *Cell* **110,** 701–711 (2002).
42. Liu, Y. *et al.* RNAi-dependent H3K27 methylation is required for heterochromatin formation and DNA elimination in *Tetrahymena*. *Genes Dev.* **21,** 1530–1545 (2007).
43. Pal-Bhadra, M. *et al.* Heterochromatic silencing and HP1 localization in *Drosophila* are dependent on the RNAi machinery. *Science* **303,** 669–672 (2004).
44. Ringrose, L. & Paro, R. Epigenetic regulation of cellular memory by the Polycomb and Trithorax group proteins. *Annu. Rev. Genet.* **38,** 413–443 (2004).
45. Deshpande, G., Calhoun, G. & Schedl, P. *Drosophila argonaute-2* is required early in embryogenesis for the assembly of centric/centromeric heterochromatin, nuclear division, nuclear migration, and germ-cell formation. *Genes Dev.* **19,** 1680–1685 (2005).
46. Haynes, K. A., Caudy, A. A., Collins, L. & Elgin, S. C. Element *1360* and RNAi components contribute to HP1-dependent silencing of a pericentric reporter. *Curr. Biol.* **16,** 2222–2227 (2006).
47. Grishok, A., Sinskey, J. L. & Sharp, P. A. Transcriptional silencing of a transgene by RNAi in the soma of *C. elegans*. *Genes Dev.* **19,** 683–696 (2005).
48. Robert, V. J., Sijen, T., van Wolfswinkel, J. & Plasterk, R. H. Chromatin and RNAi factors protect the *C. elegans* germline against repetitive sequences. *Genes Dev.* **19**, 782–787 (2005).
49. Kim, J. K. *et al.* Functional genomic analysis of RNA interference in *C. elegans*. *Science* **308,** 1164–1167 (2005).
50. Grimaud, C. *et al.* RNAi components are required for nuclear clustering of Polycomb group response elements. *Cell* **124,** 957–971 (2006).
51. Grewal, S. I. & Elgin, S. C. Transcription and RNA interference in the formation of heterochromatin. *Nature* **447,** 399–406 (2007).
52. Folco, H. D., Pidoux, A. L., Urano, T. & Allshire, R. C. Heterochromatin and RNAi are required to establish CENP-A chromatin at centromeres. *Science* **319,** 94–97 (2008).
53. Zaratiegui, M., Irvine, D. V. & Martienssen, R. A. Noncoding RNAs and gene silencing. *Cell* **128,** 763–776 (2007).
54. Grewal, S. I. & Jia, S. Heterochromatin revisited. *Nature Rev. Genet.* **8**, 35–46 (2007).
55. Rea, S. *et al.* Regulation of chromatin structure by site-specific histone H3 methyltransferases. *Nature* **406,** 593–599 (2000).
56. Lachner, M., O'Carroll, D., Rea, S., Mechtler, K. & Jenuwein, T. Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature* **410,** 116–120 (2001).

57. Bannister, A. J. *et al.* Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature* **410,** 120–124 (2001).
58. Nakayama, J., Rice, J. C., Strahl, B. D., Allis, C. D. & Grewal, S. I. Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly. *Science* **292,** 110–113 (2001).
59. Partridge, J. F., Borgstrom, B. & Allshire, R. C. Distinct protein interaction domains and protein spreading in a complex centromere. *Genes Dev.* **14,** 783–791 (2000).
60. Sadaie, M., Iida, T., Urano, T. & Nakayama, J. A chromodomain protein, Chp1, is required for the establishment of heterochromatin in fission yeast. *EMBO J.* **23,** 3825–3835 (2004).
61. Motamedi, M. R. *et al.* Two RNAi complexes, RITS and RDRC, physically interact and localize to noncoding centromeric RNAs. *Cell* **119,** 789–802 (2004).
62. Sugiyama, T., Cam, H., Verdel, A., Moazed, D. & Grewal, S. I. RNA-dependent RNA polymerase is an essential component of a self-enforcing loop coupling heterochromatin assembly to siRNA production. *Proc. Natl Acad. Sci. USA* **102,** 152–157 (2005).
These two studies identified the *S. pombe* RdRP complex and demonstrated that it has *in vitro* dsRNA synthesis activity and is associated with RITS and non-coding centromeric RNA. They further showed that the dsRNA synthesis activity is required for RNAi-mediated gene silencing.
63. Buker, S. M. *et al.* Two different Argonaute complexes are required for siRNA generation and heterochromatin assembly in fission yeast. *Nature Struct. Mol. Biol.* **14,** 200–207 (2007).
64. Irvine, D. V. *et al.* Argonaute slicing is required for heterochromatic silencing and spreading. *Science* **313,** 1134–1137 (2006).
65. Noma, K. *et al.* RITS acts *in cis* to promote RNA interference-mediated transcriptional and post-transcriptional silencing. *Nature Genet.* **36,** 1174–1180 (2004).
66. Buhler, M., Verdel, A. & Moazed, D. Tethering RITS to a nascent transcript initiates RNAi- and heterochromatin-dependent gene silencing. *Cell* **125,** 873–886 (2006).
This study demonstrated that tethering of the Tas3 component of RITS to the RNA transcript of a euchromatic reporter gene induced RNAi and heterochromatin assembly at the reporter gene.
67. Hong, E.-J. E., Villen, J., Gerace, E. L., Gygi, S. P. & Moazed, D. A Cullin E3 ubiquitin ligase complex associates with Rik1 and the Clr4 histone H3-K9 methyltransferase and is required for RNAi-mediated heterochromatin formation. *RNA Biol.* **2,** 106–111 (2005).
68. Bayne, E. H. *et al.* Splicing factors facilitate RNAi-directed silencing in fission yeast. *Science* **322,** 602–606 (2008).
69. Wutz, A. RNAs templating chromatin structure for dosage compensation in animals. *Bioessays* **25,** 434–442 (2003).
70. Iida, T., Nakayama, J. & Moazed, D. siRNA-mediated heterochromatin establishment requires HP1 and is associated with antisense transcription. *Mol. Cell* **31,** 178–189 (2008).
71. Motamedi, M. R. *et al.* HP1 proteins form distinct complexes and mediate heterochromatic gene silencing by non-overlapping mechanisms. *Mol. Cell* **32,** 778–790 (2008).
72. Colmenares, S. U., Buker, S. M., Buhler, M., Dlakic, M. & Moazed, D. Coupling of double-stranded RNA synthesis and siRNA generation in fission yeast RNAi. *Mol. Cell* **27,** 449–461 (2007).
73. Sarot, E., Payen-Groschene, G., Bucheton, A. & Pelisson, A. Evidence for a *piwi*-dependent RNA silencing of the *gypsy* endogenous retrovirus by the *Drosophila melanogaster flamenco* gene. *Genetics* **166,** 1313–1321 (2004).
74. Reiss, D., Josse, T., Anxolabehere, D. & Ronsseray, S. *aubergine* mutations in *Drosophila melanogaster* impair *P* cytotype determination by telomeric *P* elements inserted in heterochromatin. *Mol. Genet. Genomics* **272,** 336–343 (2004).
75. Chan, S. W., Zhang, X., Bernatavichute, Y. V. & Jacobsen, S. E. Two-step recruitment of RNA-directed DNA methylation to tandem repeats. *PLoS Biol.* **4,** e363 (2006).
76. Ting, A. H., Schuebel, K. E., Herman, J. G. & Baylin, S. B. Short double-stranded RNA induces transcriptional gene silencing in human cancer cells in the absence of DNA methylation. *Nature Genet.* **37,** 906–910 (2005).
77. Janowski, B. A. *et al.* Involvement of AGO1 and AGO2 in mammalian transcriptional silencing. *Nature Struct. Mol. Biol.* **13,** 787–792 (2006).
78. Sogo, J. M., Stahl, H., Koller, T. & Knippers, R. Structure of replicating simian virus 40 minichromosomes. The replication fork, core histone segregation and terminal structures. *J. Mol. Biol.* **189,** 189–204 (1986).
79. Jenuwein, T. & Allis, C. D. Translating the histone code. *Science* **293,** 1074–1080 (2001).
80. Mello, C. C. & Conte, D. Jr. Revealing the world of RNA interference. *Nature* **431,** 338–342 (2004).
81. Zhang, K., Mosch, K., Fischle, W. & Grewal, S. I. Roles of the Clr4 methyltransferase complex in nucleation, spreading and maintenance of heterochromatin. *Nature Struct. Mol. Biol.* **15,** 381–388 (2008).
82. Chen, E. S. *et al.* Cell cycle control of centromeric repeat transcription and heterochromatin assembly. *Nature* **451,** 734–737 (2008).
83. Vanacova, S. *et al.* A new yeast poly(A) polymerase complex involved in RNA quality control. *PLoS Biol.* **3,** e189 (2005).
84. Chen, C. C. *et al.* A member of the polymerase β nucleotidyltransferase superfamily is required for RNA interference in *C. elegans*. *Curr. Biol.* **15,** 378–383 (2005).
85. Lee, S. R. & Collins, K. Physical and functional coupling of RNA-dependent RNA polymerase and Dicer in the biogenesis of endogenous siRNAs. *Nature Struct. Mol. Biol.* **14,** 604–610 (2007).
86. Kloc, A., Zaratiegui, M., Nora, E. & Martienssen, R. RNA interference guides histone modification during the S phase of chromosomal replication. *Curr. Biol.* **18,** 490–495 (2008).
87. Kato, H. *et al.* RNA polymerase II is required for RNAi-dependent heterochromatin assembly. *Science* **309,** 467–469 (2005).
88. Djupedal, I. *et al.* RNA Pol II subunit Rpb7 promotes centromeric transcription and RNAi-directed chromatin silencing. *Genes Dev.* **19,** 2301–2306 (2005).
89. El-Shami, M. *et al.* Reiterated WG/GW motifs form functionally and evolutionarily conserved ARGONAUTE-binding platforms in RNAi-related components. *Genes Dev.* **21,** 2539–2544 (2007).
This study, together with references 93–95, establishes GW-motif-containing proteins as conserved Argonaute adaptors.
90. Morris, K. V., Chan, S. W., Jacobsen, S. E. & Looney, D. J. Small interfering RNA-induced transcriptional gene silencing in human cells. *Science* **305,** 1289–1292 (2004).
91. Kim, D. H., Villeneuve, L. M., Morris, K. V. & Rossi, J. J. Argonaute-1 directs siRNA-mediated transcriptional gene silencing in human cells. *Nature Struct. Mol. Biol.* **13,** 793–797 (2006).
92. Carmell, M. A. *et al.* MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline. *Dev. Cell* **12,** 503–514 (2007).
93. Eulalio, A., Huntzinger, E. & Izaurralde, E. GW182 interaction with Argonaute is essential for miRNA-mediated translational repression and mRNA decay. *Nature Struct. Mol. Biol.* **15,** 346–353 (2008).
94. Debeauchamp, J. L. *et al.* Chp1–Tas3 interaction is required to recruit RITS to fission yeast centromeres and for maintenance of centromeric heterochromatin. *Mol. Cell. Biol.* **28,** 2154–2166 (2008).
95. Till, S. *et al.* A conserved motif in Argonaute-interacting proteins mediates functional interactions through the Argonaute PIWI domain. *Nature Struct. Mol. Biol.* **14,** 897–903 (2007).
96. Li, C. F. *et al.* An ARGONAUTE4-containing nuclear processing center colocalized with Cajal bodies in *Arabidopsis thaliana*. *Cell* **126,** 93–106 (2006).
97. Pontes, O. *et al.* The *Arabidopsis* chromatin-modifying nuclear siRNA pathway involves a nucleolar RNA processing center. *Cell* **126,** 79–92 (2006).
98. Kanno, T. *et al.* Atypical RNA polymerase subunits required for RNA-directed DNA methylation. *Nature Genet.* **37,** 761–765 (2005).
99. Herr, A. J., Jensen, M. B., Dalmay, T. & Baulcombe, D. C. RNA polymerase IV directs silencing of endogenous DNA. *Science* **308,** 118–120 (2005).
100. Brower-Toland, B. *et al. Drosophila* PIWI associates with chromatin and interacts directly with HP1a. *Genes Dev.* **21,** 2300–2311 (2007).
This study identifies a physical association between HP1 and PIWI, suggesting that HP1 may collaborate with PIWI in establishing repressive chromatin domains.

Acknowledgements I thank the members of my laboratory and colleagues in the chromatin and RNA silencing fields for fruitful discussions, and the National Institutes of Health, the Leukemia and Lymphoma Society, and the Howard Hughes Medical Institute for funding.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The author declares no competing financial interests. Correspondence should be addressed to the author (danesh_moazed@hms.harvard.edu).

Viral and cellular messenger RNA targets of viral microRNAs

Bryan R. Cullen[1]

Given the propensity of viruses to co-opt cellular pathways and activities for their benefit, it is perhaps not surprising that several viruses have now been shown to reshape the cellular environment by reprogramming the host's RNA-interference machinery. In particular, microRNAs are produced by the various members of the herpesvirus family during both the latent stage of the viral life cycle and the lytic (or productive) stage. Emerging data suggest that viral microRNAs are particularly important for regulating the transition from latent to lytic replication and for attenuating antiviral immune responses.

MicroRNAs (miRNAs) are a class of small (~21–25 nucleotides) single-stranded RNAs that can inhibit the expression of specific messenger RNAs by binding to complementary target sequences within the mRNAs. Viruses were first reported to express miRNAs in 2004 (ref. 1), when Thomas Tuschl and colleagues described five miRNAs that were produced in human B cells after infection with the γ-herpesvirus Epstein–Barr virus (EBV). Subsequently, miRNAs have been found to be expressed by all of the herpesviruses examined. EBV is now known to encode at least 23 miRNAs[2,3], and the distantly related human γ-herpesvirus Kaposi's sarcoma-associated herpesvirus (KSHV) encodes 12 (refs 3–5). Similarly, the β-herpesvirus human cytomegalovirus (HCMV) encodes 11 miRNAs[5,6], and the human α-herpesvirus herpes simplex virus 1 (HSV-1) encodes at least 6 miRNAs[7,8]. (These numbers refer to the known pre-miRNA precursors encoded by each virus. A pre-miRNA can give rise to a single mature miRNA or to two miRNAs, one of which will be more abundant.) Several miRNAs have also been identified in herpesviruses that infect other species, including the simian γ-herpesviruses rhesus rhadinovirus[9] and rhesus lymphocryptovirus[2], murine γ-herpesvirus 68 (MHV68)[5], murine cytomegalovirus[10,11], and the avian α-herpesviruses Marek's disease virus types 1 and 2 (refs 12, 13).

Unlike herpesviruses (which are a family of DNA viruses), other, unrelated, DNA viruses seem to encode either one or two miRNAs (for example, primate polyoma viruses and human adenoviruses) or none at all[5,14–17]. Viruses that have an RNA genome, including retroviruses and flaviviruses, have been reported to lack miRNAs[5,17], although this result remains somewhat controversial for human immunodeficiency virus 1 (HIV-1)[18]. The absence of viral miRNAs in the RNA viruses examined so far might be partly explained by the fact that if the viral genome contained an appropriate precursor, this might be excised by the miRNA-processing enzyme of the host cell (that is, by Drosha)[19], resulting in degradation of the viral genome. Moreover, most RNA viruses — as well as DNA viruses belonging to the poxvirus family — replicate in the cytoplasm, away from the nucleus, where the Drosha-containing microprocessor complex is located. Therefore, even if the genomes of these cytoplasmic viruses encode a miRNA, it is not apparent how they could be processed to yield a mature miRNA.

It is less clear why miRNAs seem to be rare in nuclear-replicating DNA viruses that are not members of the herpesvirus family. It seems possible that the presence of miRNAs in herpesviruses is associated with the characteristic ability of herpesviruses to establish long-term latent infections. Avoiding the host immune response is particularly important during latent infection, and viral miRNAs not only have the advantage of not being recognized by the host immune system but also might be an ideal tool for attenuating immune responses by downregulating the expression of key cellular genes. Moreover, miRNAs might provide a way to regulate the entry of herpesviruses to the latent stage of the life cycle and/or their exit from this stage[20]. Other DNA virus families usually establish productive infections that often result in the infected cell's dying rapidly as a result of pathogenic factors produced by the virus or cytotoxic responses induced in the host. On the one hand, given that miRNAs operate at the level of the mRNA, they might not be as useful during a productive (lytic) replication cycle, because the proteins encoded by the targeted mRNAs might have a half-life that approaches, or even exceeds, the duration of the viral life cycle. On the other hand, viral miRNAs could be effective inhibitors of cellular mRNAs that are produced *de novo* during infection and that might encode proteins with antiviral activities. It will be interesting to see whether additional viral miRNAs, encoded by DNA viruses other than those of the herpesvirus family, will be uncovered in the future. In this Review, I briefly discuss what is known about the biogenesis and function of the known viral miRNAs, focusing on the limited number of viral and cellular mRNA targets that have been identified for these viral miRNAs so far.

Viral miRNA generation

The genomic regions encoding cellular miRNAs are generally transcribed by RNA polymerase II. The initial product is a capped, polyadenylated transcript that includes one or more stem–loop structures, each of which contains a mature miRNA sequence as part of one arm (Fig. 1). This precursor is known as a primary miRNA (pri-miRNA)[19] (see page 396 for further details about miRNA biogenesis). The nuclease Drosha cleaves the pri-miRNA stem, excising hairpin intermediates of ~65–70 nucleotides known as precursor miRNAs (pre-miRNAs). These are exported to the cytoplasm and processed by another nuclease, Dicer, generating mature miRNAs of ~22 nucleotides. The miRNAs are loaded into a protein complex known as the RNA-induced silencing complex (RISC), which they then guide to the target mRNA to exert their effector function. Binding of the RISC to an mRNA bearing extensive sequence complementarity to the miRNA generally results in mRNA cleavage and degradation, whereas binding to mRNAs bearing partial complementarity results mainly in translational arrest.

[1]Department of Molecular Genetics & Microbiology, Center for Virology, Duke University Medical Center, Durham, North Carolina 27710, USA.

At present, there is no evidence that any vertebrate virus encodes novel miRNA-processing factors or RISC components. So it seems that, in general, viral miRNAs are transcribed and processed in the same way as cellular miRNAs and, moreover, that RISCs programmed by viral miRNAs are functionally equivalent to those programmed by cellular miRNAs (Fig. 1). One exception is miRNAs encoded by MHV68 and by human adenoviruses, because the regions of the viral genome encoding these miRNAs are initially transcribed by RNA polymerase III (refs 5, 16). Although the processing steps involved in the biogenesis of MHV68 miRNAs remain to be fully elucidated, mature MHV68 and human adenovirus miRNAs are probably excised by Dicer and loaded into the RISC normally.

A key characteristic of DNA viruses is that gene expression during productive replication is temporally regulated: viral proteins can be subdivided into immediate early, early, and late species. Immediate early gene products are usually regulatory proteins. Early proteins are more diverse and are often involved in viral genome replication or in host immune response regulation. Late proteins are usually structural. The γ-herpesviruses, in particular, also produce viral proteins during latency, and these proteins have roles in episome maintenance, cell growth regulation and immune evasion. It is important therefore to consider whether the expression of viral miRNAs is also temporally regulated. For most herpesviruses, this is unclear at present, because the known viral miRNAs have been cloned either exclusively from latently infected cells or from cells at a relatively late stage in productive replication. However, there is evidence indicating that some viral miRNAs are active during latency, whereas others are more important during productive replication. For example, in the case of KSHV, all 12 viral miRNAs are derived from a cluster that is transcribed as a single pri-miRNA during latent infection. During the transition to productive replication, a viral lytic promoter is activated, and this promoter lies 3′ to the genomic sequences encoding ten of the KSHV miRNAs but 5′ to the genomic sequences encoding two of them. Consequently, the expression of ten of the KSHV miRNAs is largely unaffected when productive replication is activated, whereas the expression of two of the miRNAs is substantially induced[5,21].

In the case of HCMV, the 11 known viral miRNAs were all identified in productively infected cells[5,6], and most of these viral miRNAs seem to be produced with 'early' kinetics (that is, gene expression depends on viral immediate early transcription factors). It remains unclear which HCMV miRNAs are produced during latency. By contrast, in cells infected with HSV-1, four viral miRNAs seem to be expressed mainly during latency, one exclusively during productive replication and one during both stages[7,8]. Finally, in cells infected with simian virus 40 (SV40), which is a polyoma virus, viral miRNAs are expressed with late kinetics[14]. A full appreciation of the functions of viral miRNAs will certainly require a more detailed understanding of how their expression is regulated during the viral life cycle.

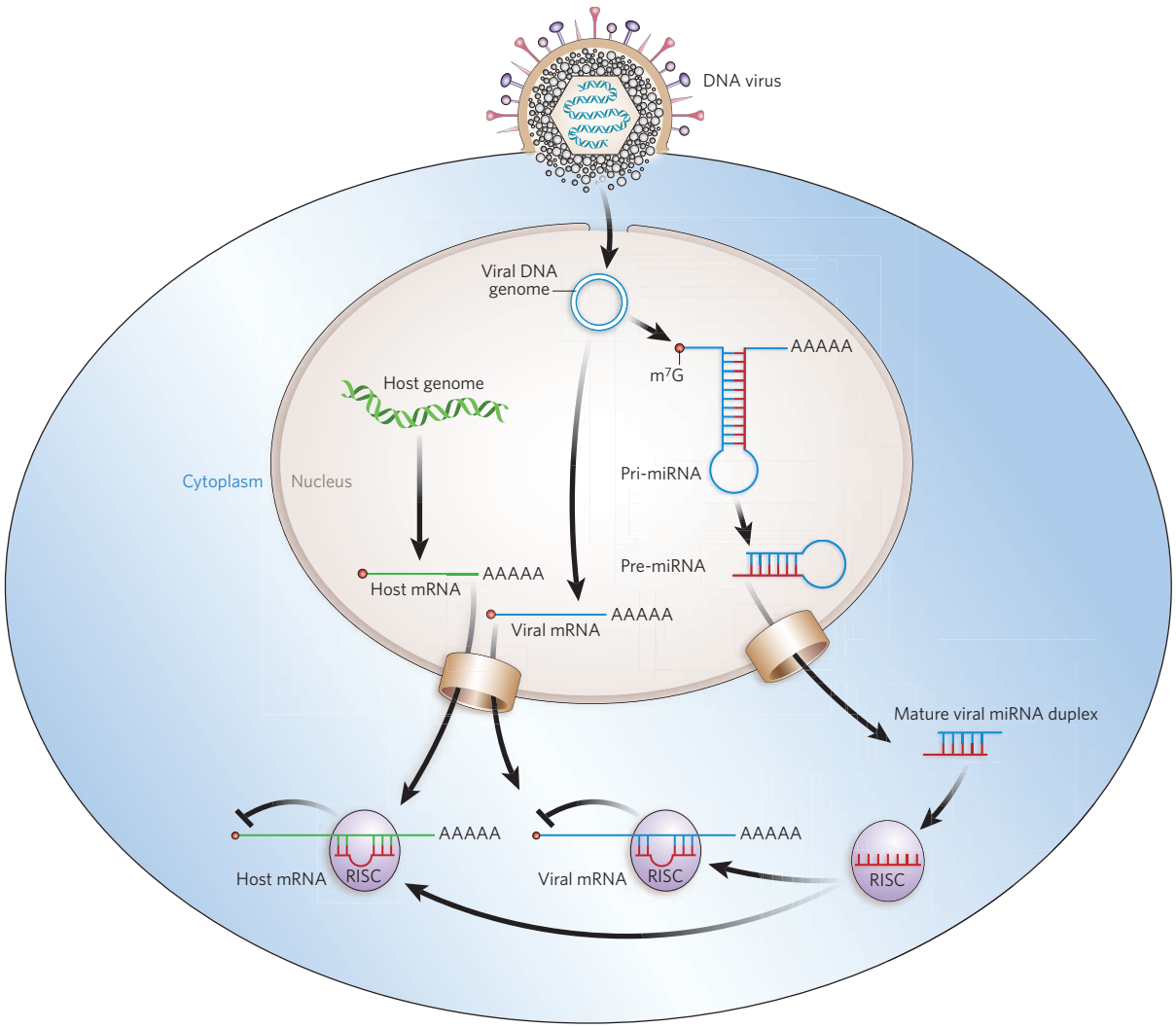


Figure 1 | How DNA virus miRNAs target host and viral mRNAs. After a host cell is infected by a DNA virus, the viral genome is transcribed in the nucleus to yield both pri-miRNAs and mRNAs. The pri-miRNA is processed by host factors in the nucleus to yield the pre-miRNA intermediate, which is then exported to the cytoplasm, where the mature viral miRNA is generated and incorporated into the RISC. RISCs that are programmed by viral miRNAs in this way can then inhibit expression of viral and/or host mRNAs in the infected cell's cytoplasm. m^7G, 7-methylguanosine.

Table 1 | Viral mRNA targets of viral miRNAs

Virus	Viral miRNA	Viral mRNA target	Function of viral protein
EBV	miR-BART2	*BALF5*	DNA polymerase
	miR-BART1-5p miR-BART16 miR-BART17-5p	*LMP1*	Signalling molecule
HvAV	miR-1	*ORF1*	DNA polymerase
SV40	miR-S1	T antigens	Early proteins
HSV-1	miR-H2-3p	*ICP0*	Immediate early protein
	miR-H6	*ICP4*	Immediate early protein
HSV-2	miR-I	*ICP34.5*	Pathogenicity factor
HCMV	miR-UL112-1	*IE1* (*IE72, UL123*)	Immediate early protein

EBV, Epstein–Barr virus; HCMV, human cytomegalovirus; HSV, herpes simplex virus; HvAV, *Heliothis virescens* ascovirus; SV40, simian virus 40.

Although many viral miRNAs have now been identified, knowledge about their functions remains scarce. There are no published reports examining the *in vivo* phenotypes of viral mutants specifically lacking individual viral miRNAs, and only a small number of mRNA targets have been described (Tables 1 and 2). It can be envisaged that viral miRNAs evolved to downregulate cellular mRNAs and/or viral mRNAs (Fig. 1). Cellular mRNA targets might include transcripts encoding proteins involved in host innate or adaptive immune responses or, more generally, involved in cell-cycle regulation or signal transduction. Viral mRNA targets might include transcripts involved in regulating the transition from latency to productive replication (or vice versa) or products of immediate early genes that need to be eliminated at later stages in the viral life cycle as a result of toxicity or because they are targets for host cytotoxic T cells[14,20]. Although the current understanding is limited, the known targets of viral miRNAs have been found to belong to almost all of these categories.

Viral mRNA targets of viral miRNAs

The first paper to describe viral miRNAs also provided the first indication of a viral mRNA target for a viral miRNA. Specifically, one of the five EBV miRNAs described by Tuschl and colleagues[1], miR-BART2, was found to lie antisense to the mRNA encoding the EBV DNA polymerase, also called *BALF5*, and was proposed to inhibit the production of DNA polymerase by inducing cleavage of this mRNA. Although there is evidence supporting partial inhibition of EBV DNA polymerase expression by miR-BART2 (ref. 22), the functional significance of this inhibition is unknown. However, inhibiting EBV DNA polymerase expression might promote entry of the virus to latency by reducing viral genome amplification early after infection. Recently, it was reported[23] that a miRNA encoded by the insect DNA virus *Heliothis virescens* ascovirus (HvAV) also downregulates expression of the viral DNA polymerase. Unlike miR-BART2, the HvAV miRNA does not lie antisense to the viral DNA polymerase mRNA and has only moderate homology to the proposed mRNA target, but a reduction in the DNA polymerase mRNA level was nevertheless observed. The fact that two miRNAs, expressed by two unrelated viral species, both reduce the level of mRNAs encoding the cognate viral DNA polymerase might indicate convergent evolution.

Another example of a viral miRNA that is transcribed antisense to a viral mRNA, and induces degradation of that mRNA, occurs in the polyoma virus SV40. SV40 encodes a single pre-miRNA stem–loop structure that is expressed exclusively as a late gene product[14]. The viral miRNAs derived from this stem–loop structure lie antisense to the early viral mRNAs that encode the SV40 T antigens, which are viral transcription factors that induce the expression of late viral genes. These SV40 miRNAs, which show perfect complementarity to the T antigen mRNAs, induce the cleavage and degradation of the mRNAs and reduce T-antigen expression late in the SV40 life cycle. Epitopes derived from SV40 T antigens are recognized by cytotoxic T cells, and the effect of these viral miRNAs is therefore to partly protect SV40-infected cells from being killed by T cells[14].

Additional cases of viral miRNAs regulating mRNAs to which they are antisense have been reported in HSV-1 and the related virus HSV-2 (refs 8, 24). During latency, HSV-1 generates a set of five miRNAs: miR-H2-3p (3p denoting derivation from the 3′ side of the pre-miRNA stem), miR-H3, miR-H4, miR-H5 and miR-H6. One of these miRNAs, miR-H2-3p, lies antisense to the mRNA encoding the viral immediate early protein ICP0 and has been shown to downregulate ICP0 production[8]. Surprisingly, miR-H2-3p does not, however, induce *ICP0* mRNA degradation, despite being fully complementary. Although the molecular basis for this phenomenon is unclear, other research groups have also reported examples of miRNAs or short interfering RNAs (a related class of small RNA) that reduce the expression of mRNAs bearing perfectly complementary targets mainly by inhibiting their translation[25,26].

In addition to miR-H2-3p lying antisense to *ICP0* transcripts, HSV-1 miR-H3 and miR-H4 lie antisense to the mRNAs encoding the pathogenicity factor ICP34.5 and, on the basis of genetic data, were proposed to inhibit ICP34.5 expression in latently infected neurons[8]. This hypothesis has now been validated for the related virus HSV-2, which encodes a miRNA similar to miR-H3 (called miR-I)[24]. When overexpressed in HSV-2-infected cells in culture, miR-I reduces the amount of *ICP34.5* mRNA expressed and the amount of protein produced. A final example of an HSV-1 miRNA that targets a viral mRNA is provided by miR-H6, which was shown to downregulate production of the HSV-1 protein ICP4 (ref. 8). This miRNA does not lie antisense to the *ICP4* gene in the HSV-1 genome, but it does show extensive complementarity to *ICP4* mRNA, including the entire miRNA sequence extending from position 2 to position 8 (the miRNA 'seed' region). Full mRNA complementarity to the miRNA seed region is generally, but not always, required for inhibition of translation[19].

Overall, it seems that four of the six known HSV-1 miRNAs function to downregulate viral mRNAs in latently infected cells. The combined action of miR-H2-3p and miR-H6, which downregulate the production of the HSV-1 immediate early proteins ICP0 and ICP4, respectively, might increase the likelihood of HSV-1 entering latency and/or inhibit the transition from latency to productive replication[8]. The inhibition of ICP34.5 expression by miR-H3 and, potentially, miR-H4 is more difficult to explain, because ICP34.5 is a pathogenicity factor that blocks activation of the host antiviral factor PKR (double-stranded-RNA-activated protein kinase) and inhibits autophagy (an innate immune response in which cells are induced to degrade the bulk of their contents, including any newly formed virion particles)[27,28]. Inhibiting ICP34.5 expression might shield latently infected neurons from the severe cytopathic effects induced by a full-blown HSV-1 productive replication cycle, and this idea is supported by the finding that HSV-1 mutants lacking the *ICP34.5* gene are much less neurotoxic[28].

Another example of a viral miRNA that downregulates a crucial viral immediate early protein is the HCMV miRNA known as miR-UL112-1, which downregulates production of the viral protein IE1 (also known as IE72 and UL123) by targeting two partly complementary sites located in the 3′ untranslated region (UTR) of *IE1* mRNAs[20,29]. This observation prompted the proposal that herpesviruses in general might use miRNAs to regulate the expression of viral proteins that can trigger the transition from latency to productive replication[6]. This hypothesis is far from proven, but two observations are consistent with the idea. First, miRNAs produced in the latent stage of HSV-1 infection downregulate production

Table 2 | Cellular mRNA targets of viral miRNAs

Virus	Viral miRNA	Host mRNA target	Function of host protein
KSHV	miR-K12-11	*BACH1* (and others)	Transcriptional suppressor
	miR-K12-6-3p (and others)	*THBS1*	Adhesion molecule, angiogenesis inhibitor
HCMV	miR-UL112-1	*MICB*	Natural-killer-cell ligand
EBV	miR-BART5	*PUMA*	Pro-apoptotic factor
	miR-BHRF1-3	*CXCL11*	Chemokine, T-cell attractant

BACH1, BTB and CNC homology 1; *CXCL11*, CXC-chemokine ligand 11; KSHV, Kaposi's sarcoma-associated herpesvirus; *MICB*, major histocompatibility complex class I polypeptide-related sequence B; *PUMA*, p53-upregulated modulator of apoptosis; *THBS1*, thrombospondin 1.

of the immediate early proteins ICP0 and ICP4 (ref. 8), as noted earlier. Second, recent data show that viral miRNAs generated during KSHV latency downregulate the production of the KSHV immediate early proteins Rta and Mta, which are known to have a key role in the activation of productive KSHV replication (P. Konstantinova and B.R.C., unpublished observations).

A final example of a viral gene product that is downregulated by viral miRNAs is the EBV protein LMP-1, which has been reported to be suppressed by three EBV miRNAs, miR-BART1-5p, miR-BART16 and miR-BART17-5p[30]. LMP-1 is a cytoplasmic signalling molecule that is produced during EBV latency and can induce cell growth and transformation. However, overexpression of LMP-1 can result in growth inhibition and increased apoptosis[30]. So the role of these miRNAs might be to ensure an optimal level of LMP-1 expression during EBV latency.

Cellular mRNA targets of viral miRNAs

In principle, viral mRNA targets for viral miRNAs should be easier to identify than cellular mRNA targets. If a viral miRNA is antisense to a viral mRNA, then this suggests an obvious potential target, although not all viral mRNAs lying antisense to a viral miRNA are downregulated by that miRNA[31]. Even if the viral miRNA interacts with a partly complementary viral mRNA, this should be an easier target to identify than a cellular mRNA, given that viral genomes are much smaller than host cell genomes. It could be envisaged that viral miRNAs evolved to efficiently degrade host cell mRNAs that encode particularly 'troublesome' host defence factors; however, no fully complementary cellular mRNA targets for viral miRNAs have been identified so far. Instead, viral miRNAs seem to inhibit the translation of cellular mRNAs bearing partly complementary sites: that is, viral miRNAs seem to function just like cellular miRNAs[19] (Table 2).

An extreme example of this is the KSHV miRNA miR-K12-11, which has a seed region identical to the human cellular miRNA miR-155 and seems to downregulate an identical, or nearly identical, set of target mRNAs[32,33]. The most fully characterized of these is *BACH1* (BTB and CNC homology 1) mRNA, which contains several targets for both miR-K12-11 and miR-155 in its 3′ UTR. BACH1 is a transcriptional suppressor, and the significance of this downregulation for KSHV replication remains unclear. Even though several human genes downregulated by both miR-K12-11 and miR-155 have been identified[32,33], it is unclear why miR-K12-11 evolved to phenocopy miR-155. Overexpression of miR-155 is, however, associated with B-cell transformation, so miR-K12-11 might contribute to the transformation of B cells by KSHV[33]. Interestingly, the avian α-herpesvirus Marek's disease virus type 1 encodes a miRNA that also functions as an orthologue of miR-155 (ref. 34), and EBV (although it does not itself encode a miR-155 equivalent) induces endogenous miR-155 production in infected B cells[35]. It therefore seems that downregulation of specific cellular genes by either miR-155 itself, or by viral orthologues of miR-155, might facilitate the replication of a range of different herpesviruses.

Another cellular gene that is downregulated by KSHV miRNAs is thrombospondin 1 (*THBS1*). *THBS1* encodes a protein that is involved in facilitating cell-to-cell adhesion and has been reported to have anti-proliferative and anti-angiogenic activities[36]. *THBS1* expression is downregulated in Kaposi's sarcoma tumours, in keeping with the fact that tumour survival, particularly in highly vascularized Kaposi's sarcoma tumours, requires angiogenesis. Rolf Renne and colleagues[36] observed that *THBS1* mRNA was downregulated in cells engineered to produce KSHV miRNAs and also showed that translation of *THBS1* mRNA is inhibited by several KSHV miRNAs, in particular by miR-K12-6-3p, which shows miRNA seed-region complementarity to two sites in the *THBS1* mRNA 3′ UTR. It therefore seems possible that downregulation of *THBS1* by KSHV miRNAs contributes to the development of Kaposi's sarcoma *in vivo*.

An obvious prediction is that viral miRNAs might downregulate cellular mRNAs encoding antiviral factors, and three such cellular targets have been uncovered. First, the HCMV miRNA miR-UL112-1 has been reported to target mRNAs encoding MICB (major histocompatibility complex class I polypeptide-related sequence B). MICB is a ligand for a cell-surface receptor of natural killer (NK) cells, which are innate immune cells that provide one of the early lines of defence against viral infection. The MICB–receptor interaction is a key regulator of NK-cell activity and hence of NK-cell killing of virus-infected cells[37]. The proposed target for miR-UL112-1 in the 3′ UTR of *MICB* mRNA is unusual in that it does not have complete complementarity to the seed region of miR-UL112-1, and in this case extensive complementarity to the central and 3′ regions of the miRNA might compensate[37]. Despite this lack of complete seed-region complementarity, cells producing miR-UL112-1 were found to display less cell-surface MICB and to be resistant to NK-cell killing *in vitro*. Conversely, cells infected with a mutant form of HCMV lacking miR-UL112-1 had more cell-surface MICB and were killed more effectively by NK cells than were cells infected with wild-type HCMV. Interestingly, the function of MICB is also inhibited by the HCMV protein UL16, suggesting that UL16 and miR-UL112-1 might be functioning synergistically to protect infected cells against the NK-cell arm of the human immune system[37]. Recently, it was reported that cellular miRNAs, including miR-93, also target the 3′ UTR of the *MICB* mRNA at sites that partly overlap with, but are distinct from, the site targeted by miR-UL112-1 (ref. 38). Although these cellular miRNAs are not similar in sequence to the viral miRNA, it seems that miR-UL112-1 is mimicking the function of a subset of cellular miRNAs and thereby exerting a similar protective effect against NK-cell killing. As discussed earlier, miR-UL112-1 has also been reported to downregulate production of the viral immediate early protein IE1 in HCMV-infected cells (Table 1), thus providing the first example of a viral miRNA that targets both viral mRNAs and cellular mRNAs.

A second example of a viral miRNA that downregulates an antiviral factor is EBV miR-BART5, which inhibits production of the pro-apoptotic protein PUMA (p53-upregulated modulator of apoptosis)[39]. Depletion of miR-BART5 from EBV-infected nasopharyngeal carcinoma cells was found to trigger higher levels of PUMA-mediated apoptosis, suggesting that miR-BART5 might shield EBV-infected epithelial cells, as well as EBV-transformed cells, from elimination by apoptosis.

The third antiviral gene product known to be downregulated by a viral miRNA is CXC-chemokine ligand 11 (CXCL11), an interferon-inducible T-cell chemoattractant. *CXCL11* mRNA is downregulated by EBV miR-BHRF1-3, which is present in large amounts in many EBV-induced B-cell tumours[40]. CXCL11 has also been shown to have anti-tumour activity in animal studies, so this finding raises the possibility that, by downregulating CXCL11 production, miR-BHRF1-3 might shield EBV-infected B cells from cytotoxic T cells *in vivo*.

Conservation of viral miRNAs

The specificity of a miRNA can be altered by changing just one or two bases, especially in the seed region[19], so the genomic sequences encoding viral miRNAs might therefore be subject to rapid evolutionary drift. But if the presence of a particular viral miRNA results in a significant increase in viral replication, then the gene encoding this miRNA might be expected to be conserved. Furthermore, if a viral miRNA targets a viral mRNA, co-evolution might be expected to occur. By contrast, if a viral miRNA targets a cellular mRNA, then the evolution of the viral miRNA gene might be expected to be restricted.

In fact, analysis of miRNAs encoded by different members of the herpesvirus and polyoma virus families has so far uncovered little sequence conservation. One exception occurs in EBV and its simian relative rhesus lymphocryptovirus: 7 of the 16 miRNAs encoded by rhesus lymphocryptovirus are markedly similar to EBV miRNAs[2]. Because EBV and rhesus lymphocryptovirus are thought to have diverged ~13 million years ago, this suggests a strong evolutionary pressure for retaining the same miRNA sequences, especially as genomic sequences adjacent to the regions encoding the mature miRNAs (for example, those encoding the terminal loop of the pre-miRNA) were found to have diverged significantly[2].

By contrast, other related viruses (for example KSHV and rhesus rhadinovirus or Marek's disease virus types 1 and 2) show no

miRNA sequence conservation[9,12], although the genomic location of the miRNAs encoded by these viruses is conserved. Conservation of genomic location, but lack of sequence similarity, is also observed for the simian polyoma virus SV40 and its human relatives the BK virus and JC virus, all of which express miRNAs that are antisense to, and degrade, viral T-antigen mRNAs[14,15]. This finding might imply that the sole function of the miRNAs expressed by these viruses is to target these particular viral mRNAs[15]. However, viral miRNAs with no known viral mRNA targets also tend to be transcribed from the same genomic location, even when their nucleotide sequences have diverged[2,9,12]. So it might simply be easier for favourable sequence changes to be selected in genomic sequences that encode a pre-existing miRNA stem–loop structure than for a novel stem–loop structure to be generated *de novo*. It therefore remains possible that these diverse polyoma virus miRNAs also target cellular mRNAs for downregulation. Moreover, the fact that two viral miRNAs have divergent sequences does not necessarily imply that they have different functions. Two distinct miRNAs, encoded by two different viruses, could, for example, target two distinct regions in a single mRNA 3′ UTR, or they could target two gene products that function at different steps in the same host metabolic pathway, resulting in a similar phenotype. Until the mRNA targets for viral miRNAs are better understood, and until there is some idea of their *in vivo* functions, the conservation (or lack of conservation) of viral miRNAs is not readily interpretable.

Outlook

Despite our still limited knowledge of viral miRNA functions, the large number of miRNAs that are encoded by diverse members of the herpesvirus family, and their high-level expression during latent infections, suggests that these small non-coding RNAs have a key role in regulating viral pathogenesis *in vivo*. In particular, it will be important to test the hypothesis that herpesvirus miRNAs that are produced during latency help to maintain the latent state[8,20], which could be examined by using viral mutants and/or antisense reagents. It certainly seems possible that antisense reagents specific for particular viral miRNAs could significantly attenuate herpesvirus-induced diseases in humans, if they could be delivered effectively to infected cells *in vivo*. ■

1. Pfeffer, S. *et al.* Identification of virus-encoded microRNAs. *Science* **304,** 734–736 (2004).
2. Cai, X. *et al.* Epstein–Barr virus microRNAs are evolutionarily conserved and differentially expressed. *PLoS Pathog.* **2,** e23 (2006).
3. Grundhoff, A., Sullivan, C. S. & Ganem, D. A combined computational and microarray-based approach identifies novel microRNAs encoded by human γ-herpesviruses. *RNA* **12,** 733–750 (2006).
4. Cai, X. *et al.* Kaposi's sarcoma-associated herpesvirus expresses an array of viral microRNAs in latently infected cells. *Proc. Natl Acad. Sci. USA* **102,** 5570–5575 (2005).
This paper showed that viral miRNAs might be conserved during viral evolution.
5. Pfeffer, S. *et al.* Identification of microRNAs of the herpesvirus family. *Nature Methods* **2,** 269–276 (2005).
This paper documented the generation of miRNAs by several herpesvirus species.
6. Grey, F. *et al.* Identification and characterization of human cytomegalovirus-encoded microRNAs. *J. Virol.* **79,** 12095–12099 (2005).
7. Cui, C. *et al.* Prediction and identification of herpes simplex virus 1-encoded microRNAs. *J. Virol.* **80,** 5499–5508 (2006).
8. Umbach, J. L. *et al.* MicroRNAs expressed by herpes simplex virus 1 during latent infection regulate viral mRNAs. *Nature* **454,** 780–783 (2008).
9. Schäfer, A., Cai, X., Bilello, J. P., Desrosiers, R. C. & Cullen, B. R. Cloning and analysis of microRNAs encoded by the primate γ-herpesvirus rhesus monkey rhadinovirus. *Virology* **364,** 21–27 (2007).
10. Buck, A. H. *et al.* Discrete clusters of virus-encoded microRNAs are associated with complementary strands of the genome and the 7.2-kilobase stable intron in murine cytomegalovirus. *J. Virol.* **81,** 13761–13770 (2007).
11. Dölken, L. *et al.* Mouse cytomegalovirus microRNAs dominate the cellular small RNA profile during lytic infection and show features of posttranscriptional regulation. *J. Virol.* **81,** 13771–13782 (2007).
12. Yao, Y. *et al.* Marek's disease virus type 2 (MDV-2)-encoded microRNAs show no sequence conservation with those encoded by MDV-1. *J. Virol.* **81,** 7164–7170 (2007).
13. Burnside, J. *et al.* Marek's disease virus encodes microRNAs that map to *meq* and the latency-associated transcript. *J. Virol.* **80,** 8778–8786 (2006).
14. Sullivan, C. S., Grundhoff, A. T., Tevethia, S., Pipas, J. M. & Ganem, D. SV40-encoded microRNAs regulate viral gene expression and reduce susceptibility to cytotoxic T cells. *Nature* **435,** 682–686 (2005).
This paper described the first viral miRNA phenotype in culture.
15. Seo, G. J., Fink, L. H., O'Hara, B., Atwood, W. J. & Sullivan, C. S. Evolutionarily conserved function of a viral microRNA. *J. Virol.* **82,** 9823–9828 (2008).
16. Xu, N., Segerman, B., Zhou, X. & Akusjarvi, G. Adenovirus virus-associated RNAII-derived small RNAs are efficiently incorporated into the RNA-induced silencing complex and associate with polyribosomes. *J. Virol.* **81,** 10540–10549 (2007).
17. Lin, J. & Cullen, B. R. Analysis of the interaction of primate retroviruses with the human RNA interference machinery. *J. Virol.* **81,** 12218–12226 (2007).
18. Ouellet, D. L. *et al.* Identification of functional microRNAs released through asymmetrical processing of HIV-1 TAR element. *Nucleic Acids Res.* **36,** 2353–2365 (2008).
19. Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116,** 281–297 (2004).
20. Murphy, E., Vanicek, J., Robins, H., Shenk, T. & Levine, A. J. Suppression of immediate-early viral gene expression by herpesvirus-coded microRNAs: implications for latency. *Proc. Natl Acad. Sci. USA* **105,** 5453–5458 (2008).
21. Gottwein, E., Cai, X. & Cullen, B. R. Expression and function of microRNAs encoded by Kaposi's sarcoma-associated herpesvirus. *Cold Spring Harb. Symp. Quant. Biol.* **71,** 357–364 (2006).
22. Barth, S. *et al.* Epstein–Barr virus-encoded microRNA miR-BART2 down-regulates the viral DNA polymerase BALF5. *Nucleic Acids Res.* **36,** 666–675 (2008).
23. Hussain, M., Taft, R. J. & Asgari, S. An insect virus-encoded microRNA regulates viral replication. *J. Virol.* **82,** 9164–9170 (2008).
24. Tang, S. *et al.* An acutely and latently expressed herpes simplex virus 2 viral microRNA inhibits expression of ICP34.5, a viral neurovirulence factor. *Proc. Natl Acad. Sci. USA* **105,** 10931–10936 (2008).
25. Wu, L., Fan, J. & Belasco, J. G. Importance of translation and nonnucleolytic Ago proteins for on-target RNA interference. *Curr. Biol.* **18,** 1327–1332 (2008).
26. Brodersen, P. *et al.* Widespread translational inhibition by plant miRNAs and siRNAs. *Science* **320,** 1185–1190 (2008).
27. He, B., Gross, M. & Roizman, B. The $\gamma_1$34.5 protein of herpes simplex virus 1 complexes with protein phosphatase 1α to dephosphorylate the α subunit of the eukaryotic translation initiation factor 2 and preclude the shutoff of protein synthesis by double-stranded RNA-activated protein kinase. *Proc. Natl Acad. Sci. USA* **94,** 843–848 (1997).
28. Orvedahl, A. *et al.* HSV-1 ICP34.5 confers neurovirulence by targeting the Beclin 1 autophagy protein. *Cell Host Microbe* **1,** 23–35 (2007).
29. Grey, F., Meyers, H., White, E. A., Spector, D. H. & Nelson, J. A human cytomegalovirus-encoded microRNA regulates expression of multiple viral genes involved in replication. *PLoS Pathog.* **3,** e163 (2007).
30. Lo, A. K. *et al.* Modulation of LMP1 protein expression by EBV-encoded microRNAs. *Proc. Natl Acad. Sci. USA* **104,** 16164–16169 (2007).
31. Grey, F. & Nelson, J. Identification and function of human cytomegalovirus microRNAs. *J. Clin. Virol.* **41,** 186–191 (2008).
32. Skalsky, R. L. *et al.* Kaposi's sarcoma-associated herpesvirus encodes an ortholog of miR-155. *J. Virol.* **81,** 12836–12845 (2007).
33. Gottwein, E. *et al.* A viral microRNA functions as an ortholog of cellular miR-155. *Nature* **450,** 1096–1099 (2007).
34. Zhao, Y. *et al.* A functional microRNA-155 ortholog encoded by the oncogenic Marek's disease virus. *J. Virol.* **83,** 489–492 (2009).
35. Yin, Q. *et al.* MicroRNA-155 is an Epstein–Barr virus-induced gene that modulates Epstein–Barr virus-regulated gene expression pathways. *J. Virol.* **82,** 5295–5306 (2008).
36. Samols, M. A. *et al.* Identification of cellular genes targeted by KSHV-encoded microRNAs. *PLoS Pathog.* **3,** e65 (2007).
37. Stern-Ginossar, N. *et al.* Host immune system gene targeting by a viral miRNA. *Science* **317,** 376–381 (2007).
38. Stern-Ginossar, N. *et al.* Human microRNAs regulate stress-induced immune responses mediated by the receptor NKG2D. *Nature Immunol.* **9,** 1065–1073 (2008).
39. Choy, E. Y. *et al.* An Epstein–Barr virus-encoded microRNA targets PUMA to promote host cell survival. *J. Exp. Med.* **205,** 2551–2560 (2008).
40. Xia, T. *et al.* EBV microRNAs in primary lymphomas and targeting of CXCL-11 by ebv-mir-BHRF1-3. *Cancer Res.* **68,** 1436–1442 (2008).

Acknowledgements Work in my laboratory was supported by the National Institutes of Health (grant numbers GM071408 and AI067968). I thank M. Luftig, E. Gottwein and J. L. Umbach for critical comments on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The author declares no competing financial interests. Correspondence should be addressed to the author (culle002@mc.duke.edu).

NATURE|Vol 457|22 January 2009|doi:10.1038/nature07758

The promises and pitfalls of RNA-interference-based therapeutics

Daniela Castanotto[1] & John J. Rossi[1]

The discovery that gene expression can be controlled by the Watson–Crick base-pairing of small RNAs with messenger RNAs containing complementary sequence — a process known as RNA interference — has markedly advanced our understanding of eukaryotic gene regulation and function. The ability of short RNA sequences to modulate gene expression has provided a powerful tool with which to study gene function and is set to revolutionize the treatment of disease. Remarkably, despite being just one decade from its discovery, the phenomenon is already being used therapeutically in human clinical trials, and biotechnology companies that focus on RNA-interference-based therapeutics are already publicly traded.

Before 1980, RNA was generally considered to be no more than a passive intermediate carrying information between DNA and protein synthesis. The discovery of catalytic RNAs in the early 1980s merited a shared Nobel prize to Tom Cech and Sidney Altman, and in 1986 the concept of 'the RNA world', an idiom created by Walter Gilbert, was proposed. Today, this is a common expression, and RNA has claimed a pivotal place in cellular biology.

Just ten years ago, RNA's functional repertoire was expanded further with the discovery in the nematode *Caenorhabditis elegans*[1] that double-stranded RNAs (dsRNAs) can trigger silencing of complementary messenger RNA sequences, and the term 'RNA interference' (RNAi) was born. Shortly thereafter, short dsRNAs — or short interfering RNAs (siRNAs) (reviewed in ref. 1) — were generated artificially and used to demonstrate that this process also occurs in mammalian cells, usually, but not always, without triggering the innate immune system, which normally recognizes RNAs as part of an antiviral defence mechanism (see page 421). The knowledge that small RNAs can affect gene expression has had a tremendous impact on basic and applied research, and RNAi is currently one of the most promising new approaches for disease therapy.

That RNAi could be triggered *in vivo* in mammals was first shown in animals infected with hepatitis B virus[2]. This was followed by the first therapeutic application of siRNAs: siRNAs were targeted to *Fas* mRNA in a mouse model of autoimmune hepatitis, resulting in protection of the treated animals against liver fibrosis[3]. In 2004, only six years after the discovery of RNAi, the first siRNA-based human therapeutics — developed as treatments for wet age-related macular degeneration — entered phase I clinical trials. RNAi is one of the fastest advancing fields in biology, and the flow of discoveries gives true meaning to the expression 'from the bench to the bedside'.

Although much is known about the mechanisms of RNAi, there are a number of challenges that applications of this gene-silencing technology need to overcome. For one, RNAi is a fundamentally important regulatory mechanism in the cell, and tapping into it in the interests of therapeutic benefit could result in side effects. Exogenously introduced dsRNA sequences can sequester components that make up the cellular machinery involved in gene silencing (see page 396), thereby reducing the accessibility of the machinery to a class of small RNAs known as microRNAs (miRNAs) that are entering the natural cellular pathway[4,5]. In addition, some synthetic siRNAs contain sequence motifs that can induce type I interferon responses and stimulate the production of pro-inflammatory cytokines[6–8].

During the past few years, many scientists have searched for solutions to overcome these limitations and to increase the safety of potential RNAi-based therapeutics. This article explores recent strategies to minimize undesirable secondary effects, describes new approaches to delivery and discusses RNAi therapies that are being tested. As it is anticipated that this technology will be applied to an increasing range of diseases, the potential problems and solutions that could one day transform RNAi into a conventional treatment for human diseases warrant careful attention.

Endogenous gene silencing

The effector RNA molecules of RNAi consist of ~20–30 nucleotides[9]. They are complexed with the protein components of the RNA-induced silencing complex (RISC). Its catalytic core in plants and animals (with the exception of single-celled organisms) is AGO2, a member of the highly conserved Argonaute protein family[10]. These small RNAs can silence gene expression by two mechanisms: post-transcriptional gene silencing (PTGS)[11], and transcriptional gene silencing (TGS)[12,13] (Fig. 1). PTGS can, in turn, be divided into two main mechanisms: direct sequence-specific cleavage, and translational repression and RNA degradation. Direct sequence-specific cleavage occurs when the targeted mRNA is perfectly complementary to the siRNA and is degraded after site-specific cleavage by the RISC. Translational repression and RNA degradation occur when the small RNA guide sequence has only limited complementarity to the target in the 'seed' region (nucleotides 2 to 8 from the 5′ end of the guide strand), with base-pairing usually occurring in the 3′ untranslated region (UTR). The latter mechanism is used by miRNAs.

TGS has been demonstrated in *Schizosaccharomyces pombe* (fission yeast), plants and, most recently, mammalian cells[14–17]. In *S. pombe*, the process is mediated by the RNA-induced transcriptional silencing complex (RITS), which contains Ago1, the chromodomain protein Chp1 and the glycine and tryptophan (GW)-repeat-containing protein Tas3 (ref. 18) (see page 413). Although in mammalian cells the mechanism by which small-RNA-directed silencing occurs is still hotly debated, both AGO1 and AGO2 have been shown to be integral to the overall

[1]Department of Molecular Biology and City of Hope Graduate School of Biological Sciences, Beckman Research Institute of the City of Hope, Duarte, California 91010, USA.

process[19,20]. Most recently, a miRNA (miR-320) has been shown to regulate transcription of the POLR3D subunit of RNA polymerase III (Pol III)[21] in human cell culture.

Endogenous small RNAs have been found in various organisms, including humans, mice, the fruitfly *Drosophila melanogaster* and *C. elegans*. Many of these originate from transposons, viruses and repetitive sequences and are characterized by their interactions with the PIWI subfamily (or PIWI clade) of Argonaute proteins[22–25] — these are thus named PIWI-interacting RNAs (piRNAs). The identification of piRNAs has been restricted to germline cells. Recently, a new class of endogenous siRNAs (endo-siRNAs or esiRNAs) has been identified in the gonads and somatic tissues of *D. melanogaster*[26–29] and in mouse oocytes[30,31]. In mice, endo-siRNAs have been proposed to regulate retrotransposon movement[30,31]. Several families of small RNAs, including repeat-associated siRNAs (ra-siRNAs), tiny non-coding RNAs (tncRNAs), *trans*-acting siRNAs (ta-siRNAs) and scan RNAs (scnRNAs) (Table 1) are found in fungi, plants and animals, but so far none of these has been observed in mammals. The evidence suggests that piRNAs act through different cellular pathways from siRNAs and miRNAs and so could offer alternative targeting strategies for therapeutic targets.

Superior designs for small molecules

Cellular genes can be targeted by exogenous introduction of siRNAs, which then take advantage of the endogenous PTGS mechanism. The siRNAs can be either transfected into cells, where they enter the RISC directly, or generated within cells through gene expression by the use of vectors containing Pol II or Pol III promoters. These RNAi triggers can be expressed in animals and plants, but not in *S. pombe*, in the form of miRNAs or as short hairpin RNAs (shRNAs), which are cleaved into small (~21–25-nucleotide) RNAs by the enzymes Drosha and/or Dicer. In both cases, if the two strands of the RNA trigger are completely complementary, the passenger strand is cleaved by AGO2 (refs 32, 33), leaving behind a single-stranded guide sequence, which acts as the template for recognition of the targeted gene sequence by the RISC (Fig. 1).

Most of the impending therapeutic applications based on RNAi propose using direct introduction of synthetic siRNAs. The advantage of

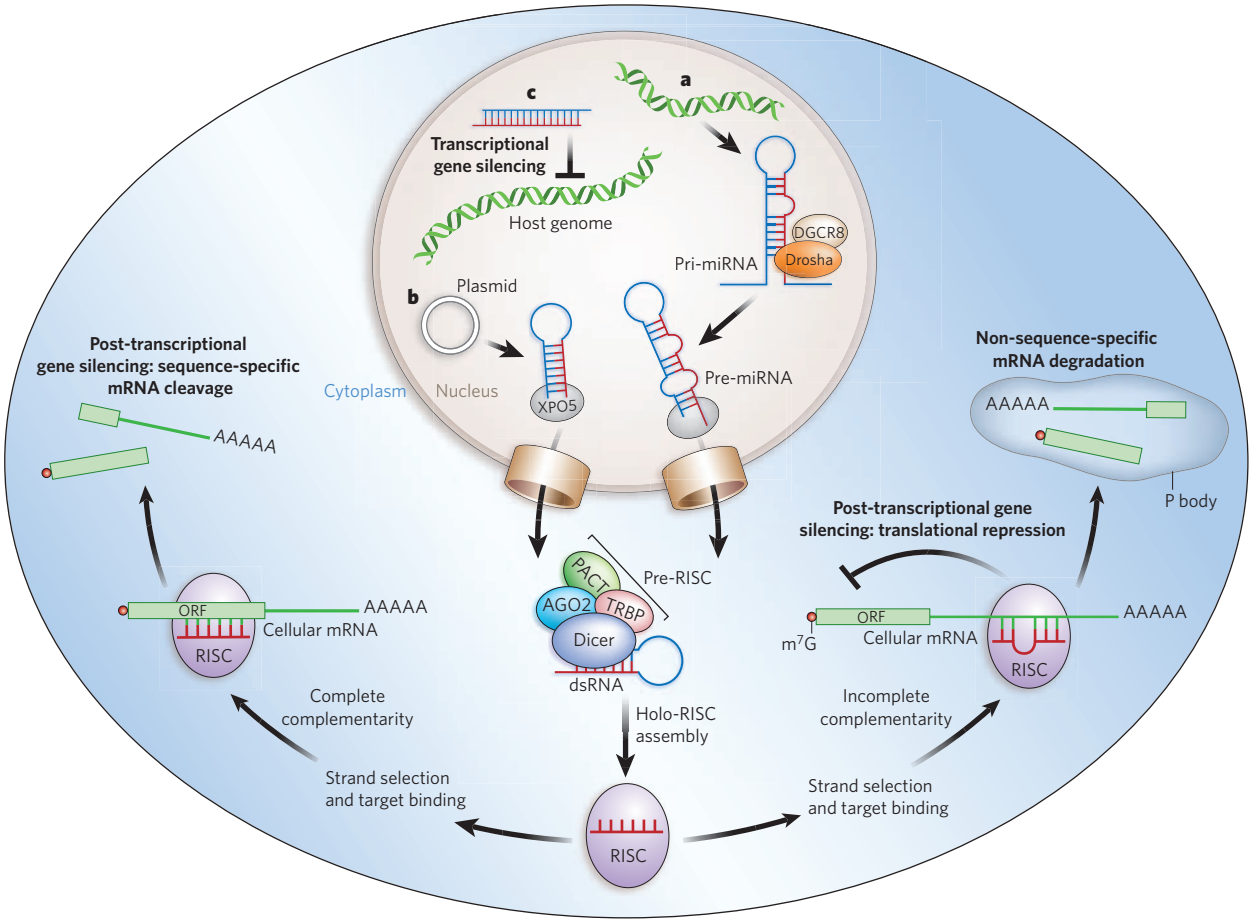


Figure 1 | Mechanisms of cellular gene silencing. **a**, Primary microRNAs (pri-miRNAs) are, in plants and animals, processed by Drosha and its partner DGCR8 into precursor miRNAs (pre-miRNAs) and then transported to the cytoplasm by exportin 5 (XPO5). In the cytoplasm, they are bound by a Dicer-containing pre-RISC and processed to yield the guide sequence that is loaded into the holo-RISC, which contains all the components required for gene silencing. AGO2 is the catalytic core of the RISC (present but not shown in the schematically drawn holo-RISC). The guide sequence binds to the corresponding target sequences in the 3′ UTRs of cellular mRNAs. If the miRNA guide sequence is fully complementary to its target site (left pathway), it triggers site-specific cleavage and degradation of the mRNA through the catalytic domain of AGO2. If the base-pairing is incomplete (right pathway) but includes pairing of the seed region (nucleotides 2–8 of the miRNA) with the target, translational inhibition occurs, and this can be accompanied by non-sequence-specific degradation of the mRNA in P bodies. **b**, Similarly to miRNAs, artificially transcribed shRNAs (in this case from a plasmid) are transported to the cytoplasm by XPO5. The dsRNA in the cytoplasm is recognized and processed by Dicer into ~21–25-nucleotide siRNA fragments that are loaded into the RISC. The siRNAs can target complementary sequences of cellular mRNAs and trigger their degradation through AGO2-mediated cleavage. **c**, When siRNAs are present in the nucleus and are complementary to promoter regions, they can trigger chromatin remodelling and histone modifications that result in transcriptional gene silencing. In mammalian cells, the details of this mechanism are still under investigation but are known to include Argonaute-family proteins. Accessory proteins indicated in the figure are TRBP (HIV *tar*-RNA-binding protein; also known as TRBP2P) and PACT (activator of protein kinase PKR; also known as PRKRA). m^7G, 7-methylguanosine.

Table 1 | Cellular small RNAs involved in gene silencing

Class	Size (nucleotides)	Functions	Mechanisms	Origin	Organisms found in
siRNAs	21–25	Regulating gene expression, providing antiviral response, restricting transposons	RNA degradation, transposon restriction	Intergenic regions, exons, introns	*Caenorhabditis elegans*, *Drosophila melanogaster*, *Schizosaccharomyces pombe*, *Arabidopsis thaliana*, *Oryza sativa* (rice)
endo-siRNAs	21–25	Restricting transposons, regulating mRNAs and heterochromatin	RNA degradation	Transposable elements, pseudogenes	*D. melanogaster*, mammals
miRNAs	21–25	Regulating gene expression at the post-transcriptional level	Blocking translation, RNA degradation	Intergenic regions, introns	*C. elegans*, *D. melanogaster*, *S. pombe*, *A. thaliana*, *O. sativa*, mammals
piRNAs	24–31*	Regulating germline development and integrity, silencing selfish DNA	Unknown	Defective transposon sequences and other repeats	*C. elegans*, *D. melanogaster*, *Danio rerio*, mammals
ra-siRNAs	23–28	Remodelling chromatin, transcriptional gene silencing	Unknown	Repeated sequence elements (subset of piRNAs)	*C. elegans*, *D. melanogaster*, *S. pombe*, *Trypanosoma brucei*, *D. rerio*, *A. thaliana*
ta-siRNAs	21–22	*Trans*-acting cleavage of endogenous mRNAs	RNA degradation	Non-coding endogenous transcripts	*D. melanogaster*, *S. pombe*, *A. thaliana*, *O. sativa*
natRNAs	21–22	Regulating gene expression at the post-transcriptional level	RNA degradation	Convergent partly overlapping transcripts	*A. thaliana*
scnRNAs	26–30	Regulating chromatin structure	DNA elimination	Meiotic micronuclei	*Tetrahymena thermophila*, *Paramecium tetraurelia*
tncRNAs	22	Unknown	Unknown	Non-coding regions	*C. elegans*

**C. elegans* piRNAs are 21 nucleotides. endo-siRNAs, endogenous siRNAs; miRNAs, microRNAs; natRNAs, natural antisense transcript siRNAs; piRNAs, PIWI-interacting RNAs; ra-siRNAs, repeat-associated siRNAs; scnRNAs, scan RNAs; siRNAs, short interfering RNAs; ta-siRNAs, *trans*-acting siRNAs; tncRNAs, tiny non-coding RNAs.

using a chemically synthesized molecule is that chemical modifications can be introduced to increase stability, promote efficacy, block binding to unintended targets that contain sequence mismatches (specific off-target effects), and reduce or abrogate potential immunostimulatory effects (general off-target effects). However, the effects of these molecules are transient, whereas the promoter-expressed shRNAs or miRNAs can potentially mediate long-term silencing with a single application.

Conventional siRNAs are ~22 nucleotides and have 3′ dinucleotide overhangs that mimic Dicer cleavage products. Because not all siRNAs achieve equivalent levels of target knockdown, large-scale siRNA screening is often performed for any given target to find the most potent inhibitors. These have yielded some rules for siRNA design. For example, to facilitate incorporation into the RISC, the 5′ end of the antisense (guide) strand should be designed to have a lower thermodynamic stability than the 5′ end of the sense strand. The proportion of the nucleotides guanosine and cytidine should be around 50% or lower, and targeting of known protein-binding sites in mRNA regulatory regions should be avoided because binding of regulatory proteins may block siRNA–target pairing. For the same reason, intramolecular structures in the target should be avoided. Statistical analyses have also found a preference for certain nucleotides at specific positions within the siRNA[34]. Many computer programs are available for identifying the optimal target sequences for a given gene[34,35]. One of these, an artificial neural network, has been used to develop a genome-wide siRNA library for humans and to identify effective siRNAs for 34 targets[36].

Chemical modifications are often included in the design of synthetic siRNAs. Selective addition of phosphorothioate linkages or substitution of 2′ fluoropyrimidines or a 2′-*O*-methyl for the 2′ ribose at certain positions does not compromise siRNA activity and concomitantly increases resistance to ribonucleases[37], which is important for *in vivo* applications. A single 2′-*O*-methyl group on the passenger strand of an siRNA duplex can abrogate activation of the Toll-like receptors[38] and prevent toxicities due to the activation of type I interferon pathway gene expression. It has recently been demonstrated that fluoro-β-D-arabinonucleic acid (FANA[39] or as 4′-*S*-FANA[40]) or arabinonucleic acid (ANA[41]) modifications can increase both the serum stability and the potency of siRNAs. Some chemical modifications also have the important advantage of decreasing or blocking the activity of the siRNA's sense (passenger) strand, thereby reducing specific off-target effects. Other modifications, such as the addition of lauric acid, lithocholic acids and cholesterol derivatives, can be made to increase cellular uptake[42], which is currently one of the main hurdles of RNAi therapy.

Breaking and entering

Therapeutic applications of siRNAs require effective delivery to the target cells and tissues. The two main strategies are delivery of chemically synthesized siRNAs (non-viral delivery), or delivery of shRNA-encoding genes by engineered viruses that will ultimately generate siRNAs by transcription in the target cells.

Non-viral delivery

Because of their size and negative charge, siRNAs cannot easily cross cell membranes. Delivery has therefore been one of the major challenges for RNAi technology. Various means of delivery have been developed and tested in murine and non-human primate models, ranging from the injection of naked RNAs into a target organ such as the lung or eye to systemic delivery of the RNA in nanoparticles, complexed with polycations, attached to cholesterol groups or conjugated with cell-surface receptors. Some delivery approaches are detailed in Fig. 2.

Two polymers that have been examined for their delivery properties are atelocollagen and chitosan. Chitosans have mucoadhesive properties and have been used for intranasal delivery to bronchiolar epithelial cells[43]. Intranasal delivery has proved an effective means of delivering siRNA in mice[44] and in non-human primates[45] to block respiratory syncytial virus infection of the upper respiratory tract. In fact, the delivery of siRNAs to mucosal membranes seems to be an effective approach in general. For example, intravaginal delivery of lipid-encapsulated siRNAs targeting herpes simplex virus 2 (HSV-2) provided protection against lethal viral infection in more than two-thirds of the siRNA-treated mice[46].

Targeting of anti-apolipoprotein B (APOB) and peroxisome proliferator-activated receptor-α (PPAR-α) siRNAs to the liver has been achieved by means of a 'membrane-active' polymer, which can mask its activity until it reaches the endosome, resulting in the delivery of siRNAs to hepatocytes after a simple intravenous injection[47]. A different siRNA delivery approach used transferrin conjugated to a cyclodextrin-polycation polymer to deliver siRNAs targeting the Ewing's sarcoma *Ews–Fli1* fusion mRNA by means of the transferrin receptor in mice[48], resulting in inhibition of tumour progression. And conjugation of an siRNA to a cholesterol group permitted its delivery to the liver and the jejunum, where it downregulated its target, APOB, leading to consequent lowering of blood cholesterol levels in a murine model system[49].

An important advance for siRNA delivery was the successful application of stable nucleic-acid lipid particles decorated with polyethylene glycol (PEG) polymer chains (termed SNALPs) for the delivery of siRNAs directed against *APOB* mRNA (*APOB*-targeted siRNAs) to the livers of

non-human primates[50]. In this case, the siRNA effect of a single intravenous injection lasted for more than 11 days and resulted in greater than 90% target knockdown and no toxicity[50]. These exciting results have increased confidence in the potential of therapeutic siRNAs for treating liver diseases.

Until recently, most approaches to *in vivo* delivery have targeted a particular organ, primarily the eye, the lungs or the liver. A significant advance in targeting siRNAs to a specific class of leukocytes involved in gut inflammation has now been reported[51]. In this study, a cyclin D1 (*Cyd1*)-targeted siRNA was loaded into stabilized nanoparticles, the surfaces of which incorporated an antibody specific for a receptor expressed by the leukocytes. The targeted siRNA-containing nanoparticles downregulated the cyclin D1 target, suppressed leukocyte proliferation and reversed experimentally induced colitis in mice[51].

Delivery of siRNAs to the nervous system has been particularly challenging. The brain is notoriously refractory to targeting because of difficulties in crossing the blood–brain barrier. However, delivery of siRNAs to the peripheral nervous system by direct infusion into the brain for the relief of chronic pain[52–54] or anxiety[55] has been demonstrated in rats. Conjugates of liposomes and antibodies or neuropeptides have also been used to deliver siRNAs into the murine brain[56]. Nevertheless, these methods do not target neurons, and a less invasive alternative to direct cranial injection is required to make such therapies more palatable.

A recent study unlocked the possibility of selective delivery of siRNAs to the central nervous system by systemic intravenous injection[57]. The siRNA involved — designed to target Japanese encephalitis virus — was conjugated with a short peptide derived from the rabies virus glycoprotein, which binds to the neuronal cell acetylcholine receptor. After transvascular delivery, 80% of the mice treated with the therapeutic siRNA survived infection with Japanese encephalitis virus, whereas 100% of the untreated controls died from complications of the infection[57].

Another interesting approach that allows systemic and targeted siRNA delivery uses a protamine–antibody fusion protein[58]. The protamine moiety is linked to the heavy-chain antigen-binding region (Fab) of an antibody to the human immunodeficiency virus 1 (HIV-1) envelope protein gp160. The positively charged protamine binds the negatively charged siRNAs — which are targeted against the HIV gene *gag* — allowing selective delivery to cells expressing the gp160 envelope protein on their surfaces[58]. This results in internalization of the antibody–siRNA complex, release of the siRNAs and downregulation of the HIV Gag-encoding

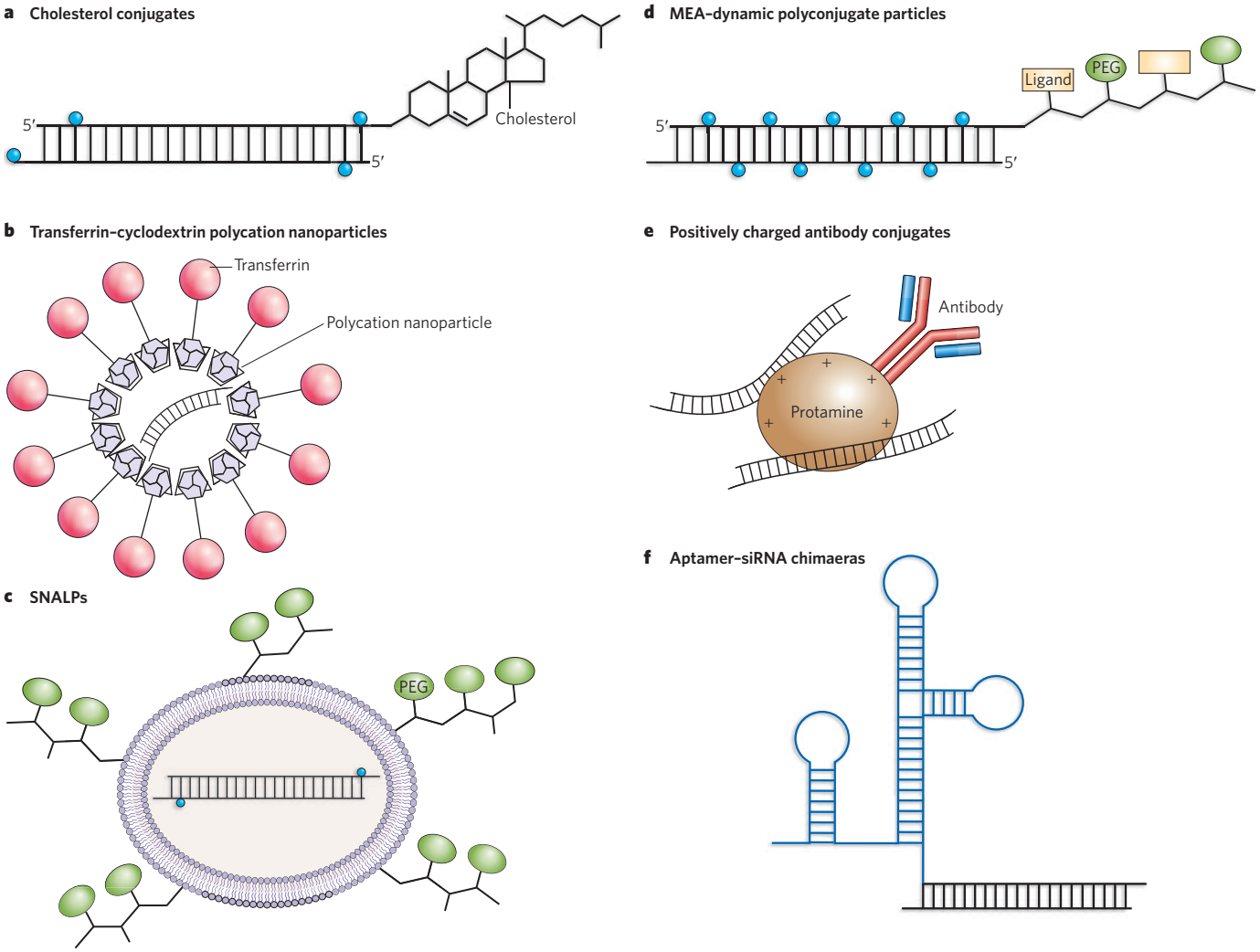


Figure 2 | *In vivo* delivery strategies for therapeutic siRNAs. **a**, Cholesterol groups can be linked to modified siRNAs to enhance their stability before systemic delivery. The most common siRNA modifications are 2′-*O*-methyluridine or 2′-fluorouridine substitutions (blue circles) combined with phosphorothioate linkages. **b**, Polycation nanoparticles can direct delivery of the siRNAs to specific cells through the use of surface ligands (such as transferrin) that bind to receptors on target cells. **c**, SNALPs encapsulate modified siRNAs into cationic or neutral lipid bilayers coated with diffusible PEG–lipid conjugates. SNALPs allow siRNAs to be taken up by cells and released by endosomes. **d**, Masked endosomolytic agent (MEA)–Dynamic PolyConjugates (DPCs) are similar to SNALPS but smaller, and contain a ligand that allows targeted cell delivery. The release of the siRNA from the endosome is also improved by the inclusion of a pH-labile bond in the MEA–DPC particles. **e**, Tagging specific antibodies with protamine or other positive charges allows the delivery of siRNAs to specific cell types via receptor-mediated uptake. **f**, Chemically linking or co-transcribing siRNAs with RNA aptamers allows the targeted delivery of the siRNAs to cells expressing the appropriate receptor.

Table 2 | Current clinical trials of RNAi-based therapeutics

siRNA	Company	Disease	Stage
Bevasiranib	Acuity Pharmaceuticals	Wet age-related macular degeneration	Phase III
		Diabetic macular oedema	Phase II
Sirna-027	Merck–Sirna Therapeutics	Wet age-related macular degeneration	Phase II
RTP801i-14	Quark Pharmaceuticals, and Silence Therapeutics	Wet age-related macular degeneration	Phase I/IIA
ALN-RSV01	Alnylam Pharmaceuticals	Respiratory syncytial virus infection	Phase II
NUC B1000	Nucleonics	Hepatitis B	Phase I
Anti-*tat/rev* shRNA	City of Hope National Medical Center, and Benitec	AIDS	Pilot feasibility study
CALAA-01	Calando Pharmaceuticals	Solid tumours	Phase I
TD101	TransDerm, and the International Pachyonychia Congenita Consortium	Pachyonychia congenita	Phase I

transcripts in a murine model *in vivo*. In this same study, fusion to protamine of an antibody specific for the hormone receptor ERBB2 allowed siRNA targeting of cancer cells expressing that receptor[58].

A similar technology for specific targeted delivery is based on aptamer–RNAi chimaeras[59]. Aptamers are *in vitro*-evolved, synthetically prepared nucleic acids that selectively bind specific ligands. An RNA aptamer designed to bind prostate-specific membrane antigen (PSMA; also known as FOLH1) was linked to a *PLK1*-targeted siRNA, and binding of the aptamer to the PSMA receptor resulted in the selective delivery into prostate cancer cells of siRNAs that target pro-survival genes[59,60]. Intratumoral injection of the PSMA–*Plk1*-targeted siRNA or PSMA–*Bcl2*-targeted siRNA conjugates into a mouse xenograft model resulted in triggering of apoptosis, growth inhibition and tumour regression[59].

A different conjugation of an siRNA to vitamin-A-coupled liposomes succeeded in delivering antifibrotic siRNAs to hepatic stellate cells, which are produced in response to liver damage[61]. In this study, multiple siRNA treatments targeting collagen chaperone-encoding genes reversed liver fibrosis by preventing collagen deposition and increased survival in rats, providing a potential therapeutic approach to treating liver cirrhosis.

Also noteworthy is the recent report of libraries of lipid-like molecules (lipidoids) that can be selected for siRNA delivery to various tissues[62].

Viral delivery

An alternative means of triggering RNAi is through promoter-expressed siRNA sequences processed from shRNAs or miRNA mimics. The genes encoding these hairpin structures are most commonly inserted into the backbones of viral vectors under the control of Pol II or Pol III promoters. A potential advantage of vector delivery is that a single administration triggers long-term expression of the therapeutic RNAi. This is particularly appropriate for chronic viral diseases such as HIV and viral hepatitis.

Lentiviral vectors have been used successfully to deliver shRNA constructs in various mammalian systems. For example, it was shown that downregulation of an activated *Ras* oncogene by a lentiviral-delivered shRNA resulted in suppression of tumour growth in mice[63]. And downregulation of the expression of a mutant form of superoxide dismutase 1 (SOD1) in mouse models of amyotrophic lateral sclerosis delayed the onset of disease[64,65]. More recently, a lentiviral vector was used to deliver a *Smad3*-targeted shRNA for regeneration of satellite cells and repair of old tissue in aged and injured muscle[66]. Viral-vector expression of shRNAs has also been explored in mouse models of neurodegenerative disorders such as Huntington's disease and Alzheimer's disease[67].

To deliver genes to the central nervous system, adenoviral vectors have proved very useful. For instance, direct brain injection of an adenoviral vector expressing a shRNA directed against the mRNA encoding the polyQ-harbouring SCA1-encoding transcript of spinocerebellar ataxia type 1 was shown to be an effective treatment in a mouse model of this disorder[68].

Despite the successes of viral delivery, it is important to bear in mind that although some viruses are non-pathogenic, they are still potentially immunogenic. Another major concern with this technique is the risk of incurring mutations in viral sequences, causing insertional mutagenesis or triggering aberrant gene expression. However, viral vectors can transduce both dividing and non-dividing cells, yield a prolonged expression of the therapeutic gene and need not be delivered in large doses. Ultimately, any therapeutic gene when expressed in large quantities has the potential to cause toxicity and immunogenicity. Critical parameters such as tolerability, long-term expression, efficacy and the ability to regulate expression and targeting should be taken into consideration when choosing a delivery method. There is no ideal delivery system for every application; rather, the delivery method needs to be tailored to the application.

Clinical trials using RNAi to treat human diseases

For a new technology, siRNAs have moved into the clinic at an unprecedented pace. Some examples of the diseases and siRNA-targeting strategies that are currently under investigation are described below.

The first siRNA protocol granted investigational new drug (IND) status and tested in a human clinical trial is the vascular endothelial growth factor (*VEGF*)-targeted siRNA Bevasiranib (Acuity Pharmaceuticals, Philadelphia, Pennsylvania) for the treatment of wet age-related macular degeneration (see Table 2 for a summary of ongoing siRNA clinical trials). This involves the overgrowth of blood vessels behind the retina, and causes severe and irreversible loss of vision; it affects 1.6 million people in the United States alone, and it is predicted that 11 million individuals worldwide will have the disease by 2013. Preclinical studies of Bevasiranib in mice showed reduced neovascularization resulting from downregulation of *Vegf* expression after direct ocular injection of the siRNA[69]. This siRNA, which is now in a phase III trial, is also in a phase II clinical trial for the treatment of diabetic macular oedema. By the conclusion of these trials, several hundred patients will have received the siRNA treatments.

Two other companies are also focusing on siRNA-based treatments against macular degeneration: Merck's Sirna Therapeutics (San Francisco, California) with an siRNA (Sirna-027) that targets the VEGF receptor VEGFR1, and Quark Pharmaceuticals (Fremont, California) in collaboration with Silence Therapeutics (London and Berlin; previously SR Pharma), with one targeted against a hypoxia-inducible gene, *RTP801* (also known as *DDIT4*), that is known to be involved in disease progression. This siRNA, RTP801i-14, has been licensed to Pfizer, UK, which is now running a phase I/IIA clinical trial. Quark Pharmaceuticals has also received IND status for another preclinical trial, in which it is currently enrolling patients. This trial is for an siRNA targeting *TP53* mRNA (which encodes the protein p53), inhibition of which delays the induction of cell-death pathways and thereby reduces acute kidney injury after surgery.

Calando Pharmaceuticals (Pasadena, California), meanwhile, has initiated a phase I clinical trial for solid tumours using an siRNA that targets a subunit of ribonucleotide reductase (RRM2), an enzyme required for the synthesis of DNA building blocks. Importantly, this trial is the first to utilize receptor-mediated delivery of siRNAs, which are encapsulated in cyclodextrin particles decorated with transferrin. This results in uptake by cells expressing the transferrin receptor, which is highly expressed on cancer cell surfaces.

The clinical trials performed by Acuity Pharmaceuticals and Merck's Sirna Therapeutics successfully stabilized patients' conditions against further degeneration and improved their vision without adverse effects. These results engendered great optimism for intravitreal injection of siRNAs, but in a stunning turn of events a report by Kleinman *et al.* demonstrated that the observed decrease in vascularization could be a consequence not of an siRNA-specific effect on angiogenesis, but rather a nonspecific activation of Toll-like receptor 3 (TLR3) and subsequent activation of interferon-γ and interleukin 12, which, in turn, downregulate VEGF[70]. In other words, both the targeted and the control siRNAs mediated nonspecific inhibition of angiogenesis through a direct interaction of the siRNAs with TLR3. Cellular uptake is not necessary for this effect, and because TLR3 is involved in several other cellular pathways the finding has highlighted another level of concern for safe clinical use of siRNAs.

Alnylam Pharmaceuticals (Cambridge, Massachusetts) is a well-established siRNA-therapeutics company whose leading candidate siRNA, ALN-RSV01, is now in a phase II clinical trial. This siRNA targets respiratory syncytial virus — which affects almost 300,000 people every year in the United States alone — by silencing the virus's nucleocapsid 'N' gene, a gene essential to viral replication. ALN-RSV01 was the first antiviral siRNA to enter clinical trials, and trials will soon be expanded to paediatric patients. Thus far it has been shown to be effective and well tolerated. Recently, Alnylam Pharmaceuticals formed an exclusive alliance with Kyowa Hakko Kogyo to develop and commercialize ALN-RSV01 in Japan and other Asian countries.

Also in development at Alnylam Pharmaceuticals are siRNAs directed against genes implicated in hypercholesterolaemia, Huntington's disease (in a joint venture with Medtronic of Minneapolis, Minnesota), hepatitis C (in a joint venture with Isis Pharmaceuticals in Carlsbad, California), progressive multifocal leukoencephalopathy (in a joint venture with Biogen Idec of Cambridge, Massachusetts) and pandemic flu (in a joint venture with the Swiss company Novartis).

The International Pachyonychia Congenita Consortium (IPCC), in collaboration with TransDerm (Santa Cruz, California), has developed an siRNA to allow the correct production of keratin as a treatment for a rare skin disorder called pachyonychia congenita.

The City of Hope National Medical Center in Duarte, California, in collaboration with Benitec (Melbourne, Australia), has started a phase I trial for the treatment of AIDS lymphoma. This trial uses a Pol III promoter-expressed shRNA targeting the HIV *tat* and *rev* shared exons. The shRNA has been incorporated into an HIV-based lentiviral vector, which in turn has been used to insert the shRNA gene (along with two other RNA-based anti-HIV genes) into blood stem cells[71]. The gene-modified stem cells have been infused into HIV-positive patients in a trial that uses autologous bone marrow transplantation to treat AIDS-related lymphomas. Four patients have now been treated in this trial.

As indicated above, partnerships have become quite accepted in the field of siRNA biotechnology. These consortia are considerably increasing the capital available for these efforts and are shortening the time involved in commercializing siRNA-based drugs.

Some companies, such as Regulus Therapeutics (Carlsbad, California), have chosen to focus on miRNAs as therapeutic targets. Santaris Pharma in Hørsholm, Denmark, has recently started the first phase I trial to target a human miRNA (miR-122). In this trial, miR-122 is being targeted for downregulation with a locked nucleic acid (LNA) anti-miRNA (SPC3649). LNA is a backbone modification that enhances the hybridization of the oligonucleotide with its target and protects it from nuclease degradation. The approach is intended to treat hepatitis C virus infection because miR-122 facilitates replication of this virus in the liver[72,73]. Downregulation of miR-122 is also potentially useful in the treatment of hypercholesterolaemia. Targeting miRNAs expressed in the heart, such as miR-208, which regulates cardiac hypertrophy and fibrosis[74], may have an advantage, because in the medical field there is a considerable experience in delivering drugs directly into this organ.

Gain or loss of miRNA function has been linked to the onset and progression of various diseases[75–77]. Protein function can be regulated either directly or indirectly by miRNAs, and alterations in miRNA expression can have profound effects on gene regulation. In instances in which disease results from altered miRNA expression, it is conceivable that normal levels could be achieved, either by targeting the specific miRNA if expression is too high or by delivering a miRNA mimic if expression is too low. However, the specificity and efficacy of delivery systems would need to be improved for this goal to be accomplished. Moreover, correct modulation of the targeted miRNA's expression is not an easy task, and it is not clear whether one miRNA can be specifically targeted without affecting other miRNAs of the same family.

The regulatory complexities of miRNAs should also be taken into consideration when either ablation or restoration of miRNA function is being considered in a therapeutic setting. A single miRNA can regulate the levels of hundreds of proteins[78,79], raising cautionary flags about the consequences of downregulating or ectopically expressing even a single miRNA species.

The safety issue

The application of siRNAs to therapeutics has raised a number of concerns about their safety. After the initial excitement, a number of reports underscored potential drawbacks to this promising technology. The first warning came from a study that recorded the deaths of mice after Pol III promoter-driven expression of shRNAs in the liver[4]. The exact mechanisms leading to mortality are still under investigation, but seem to be due at least in part to saturation of the transport factor, exportin 5, that ferries miRNAs from the nucleus to the cytoplasm. There are now indications that other factors involved in the RNAi process can also be saturated by high-level expression of exogenous siRNAs, which can sequester them from their cognate cellular miRNAs. Because each cellular miRNA can potentially modulate the expression of several hundred genes[78,79], minor alterations in the miRNA pathway can have major consequences.

One strategy to mitigate this problem is to use the lowest possible concentration of siRNAs that provides therapeutic efficacy by designing the exogenous siRNAs to be Dicer substrates (by increasing their length). These RNAs enter the RNAi pathway upstream of the RISC at the step of Dicer cleavage, which facilitates passing the siRNA to AGO2 for selection of the guide strand, often resulting in enhanced RNAi at lower concentrations than can be achieved with the exogenous delivery of cognate 21-base siRNAs[80–83]. Although small amounts of siRNAs are not expected to saturate the RNAi machinery, they can compete with miRNAs for selective incorporation into the RISC[5]. The long-term consequences of such competition are poorly understood.

With the use of microarrays, it has become increasingly obvious that introducing foreign siRNAs into the cell alters the expression of non-target genes, as well as target genes[84,85]; as few as six or seven nucleotides complementary to the seed region could result in a specific off-target effect[86] through a miRNA-like mechanism. Because microarrays only reflect mRNA levels, they do not take into account any genes affected at the translational level, and so at present it is not clear how extensive the problem of off-target effects really is. Given that the application of synthetic siRNAs results in transient inhibition of gene expression, specific off-targeting may not be a major concern for many clinical applications. Nevertheless, appropriate toxicity testing should take into account the potential for a particular siRNA to target 3′ UTRs in non-target genes.

Some strategies can be used in siRNA design to minimize the problem of off-targeting. For instance, it has been shown that 2′-*O*-Me modifications[87] or DNA substitutions[88] in siRNA duplexes can significantly reduce off-target effects. It would also be valuable to improve antisense-strand selectivity by taking into account thermodynamic stability (see 'Superior designs for small molecules') or by blocking the 5′ phosphorylation of the sense strand[89].

RNAi is a widely conserved mechanism that may originally have evolved to combat viral infections. As such, it is perhaps not surprising that in some cases siRNAs can act as agonists of Toll-like receptors[90] and that specific sequence motifs, such as uridine-rich regions and guanosine- and uridine-rich regions, can induce cellular immune responses[6,7].

The ability of an siRNA to stimulate cellular immune responses is based not only on specific sequences but also on structure, the type of delivery system used and the cell type[7,91]. Although the immunostimulatory potential of siRNAs could be advantageous in certain circumstances[92], it is usually an unwanted outcome. The above-mentioned finding of the TLR3 response to non-sequence-specific modulation of VEGF or the VEGF receptor[70], as well as a separate report showing that a macrophage migration inhibitory factor (*Mif*)-targeted siRNA (in a murine model) and a nonspecific control siRNA increased the proliferation of breast cancer cells through activation of dsRNA-activated protein kinase (PKR)[93], raise serious concerns in interpreting the results of *in vivo* siRNA applications.

Although we have yet to reach a universal solution for avoiding all off-target effects, it is foreseeable that these problems will be overcome by the use of appropriate backbone modifications, as well as delivery systems that can mask RNAs from the receptors of the innate immune system[94].

Gazing ahead

Despite the technique's youth, the list of diseases for which RNAi is being tested as a therapeutic agent is extensive, and includes Parkinson's disease, Lou Gehrig's disease, HIV infection, wet age-related macular degeneration, type 2 diabetes, obesity, hypercholesterolaemia, rheumatoid arthritis, respiratory diseases and cancers. It is already a multimillion dollar business, projected to reach US$1 billion by 2010, and intellectual property rights will become an increasingly important concern in the coming years.

However, although much has been accomplished, obstacles remain that will hamper the race to the clinic. The ultimate goal of achieving RNAi-based therapies for life-threatening or debilitating diseases cannot be attained without improving the safety, effectiveness and reliability of RNAi-trigger delivery systems. The use of targeted delivery strategies that permit systemic delivery will be a big step towards fulfilling this difficult task. The development of new, noninvasive imaging methods to monitor the *in vivo* delivery of siRNAs, such as labelling with near-infrared dyes[95], will aid studies of tissue uptake and biodistribution.

Although RNAi is not yet an accepted therapeutic modality, the enormous interest in this phenomenon ensures that we will soon witness fast advances and new applications for RNAi-based therapies. Given the way that RNAi has transformed basic research and the unprecedented speed with which it has reached the clinic, the coming years promise to be exciting. ■

1. Zamore, P. D. RNA interference: big applause for silencing in Stockholm. *Cell* **127,** 1083–1086 (2006).
2. McCaffrey, A. P. *et al.* RNA interference in adult mice. *Nature* **418,** 38–39 (2002).
This study was the first to show siRNA activity *in vivo* in mammals.
3. Song, E. *et al.* RNA interference targeting Fas protects mice from fulminant hepatitis. *Nature Med.* **9,** 347–351 (2003).
This paper provided the first therapeutic RNAi demonstration in animals.
4. Grimm, D. *et al.* Fatality in mice due to oversaturation of cellular microRNA/short hairpin RNA pathways. *Nature* **441,** 537–541 (2006).
This article raised cautionary concerns about the danger of high-level shRNA expression in animals.
5. Castanotto, D. *et al.* Combinatorial delivery of small interfering RNAs reduces RNAi efficacy by selective incorporation into RISC. *Nucleic Acids Res.* **35,** 5154–5164 (2007).
6. Hornung, V. *et al.* Sequence-specific potent induction of IFN-α by short interfering RNA in plasmacytoid dendritic cells through TLR7. *Nature Med.* **11,** 263–270 (2005).
7. Judge, A. D. *et al.* Sequence-dependent stimulation of the mammalian innate immune response by synthetic siRNA. *Nature Biotechnol.* **23,** 457–462 (2005).
8. Gantier, M. P. & Williams, B. R. The response of mammalian cells to double-stranded RNA. *Cytokine Growth Factor Rev.* **18,** 363–371 (2007).
9. Elbashir, S. M. *et al.* Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature* **411,** 494–498 (2001).
This study was the first to show that RNAi triggers can work in mammalian cells without stimulating interferon pathways.
10. Tolia, N. H. & Joshua-Tor, L. Slicer and the argonautes. *Nature Chem. Biol.* **3,** 36–43 (2007).
11. Zamore, P. D., Tuschl, T., Sharp, P. A. & Bartel, D. P. RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell* **101,** 25–33 (2000).
12. Matzke, M. A. & Birchler, J. A. RNAi-mediated pathways in the nucleus. *Nature Rev. Genet.* **6,** 24–35 (2005).
13. Wassenegger, M. The role of the RNAi machinery in heterochromatin formation. *Cell* **122,** 13–16 (2005).
14. Wassenegger, M., Heimes, S., Riedel, L. & Sanger, H. L. RNA-directed *de novo* methylation of genomic sequences in plants. *Cell* **76,** 567–576 (1994).
15. Morris, K. V., Chan, S. W., Jacobsen, S. E. & Looney, D. J. Small interfering RNA-induced transcriptional gene silencing in human cells. *Science* **305,** 1289–1292 (2004).
16. Castanotto, D. *et al.* Short hairpin RNA-directed cytosine (CpG) methylation of the *RASSF1A* gene promoter in HeLa cells. *Mol. Ther.* **12,** 179–183 (2005).
17. Ting, A. H., Schuebel, K. E., Herman, J. G. & Baylin, S. B. Short double-stranded RNA induces transcriptional gene silencing in human cancer cells in the absence of DNA methylation. *Nature Genet.* **37,** 906–910 (2005).
18. Verdel, A. *et al.* RNAi-mediated targeting of heterochromatin by the RITS complex. *Science* **303,** 672–676 (2004).
19. Janowski, B. A. *et al.* Involvement of AGO1 and AGO2 in mammalian transcriptional silencing. *Nature Struct. Mol. Biol.* **13,** 787–792 (2006).
20. Kim, D. H., Villeneuve, L. M., Morris, K. V. & Rossi, J. J. Argonaute-1 directs siRNA-mediated transcriptional gene silencing in human cells. *Nature Struct. Mol. Biol.* **13,** 793–797 (2006).
21. Kim, D. H., Saetrom, P., Snove, O. Jr & Rossi, J. J. MicroRNA-directed transcriptional gene silencing in mammalian cells. *Proc. Natl Acad. Sci. USA* **105,** 16230–16235 (2008).
22. Aravin, A. A., Hannon, G. J. & Brennecke, J. The Piwi–piRNA pathway provides an adaptive defense in the transposon arms race. *Science* **318,** 761–764 (2007).
23. Klattenhoff, C. & Theurkauf, W. Biogenesis and germline functions of piRNAs. *Development* **135,** 3–9 (2008).
24. Batista, P. J. *et al.* PRG-1 and 21U-RNAs interact to form the piRNA complex required for fertility in *C. elegans*. *Mol. Cell* **31,** 67–78 (2008).
25. Das, P. P. *et al.* Piwi and piRNAs act upstream of an endogenous siRNA pathway to suppress Tc3 transposon mobility in the *Caenorhabditis elegans* germline. *Mol. Cell* **31,** 79–90 (2008).
26. Ghildiyal, M. *et al.* Endogenous siRNAs derived from transposons and mRNAs in *Drosophila* somatic cells. *Science* **320,** 1077–1081 (2008).
27. Czech, B. *et al.* An endogenous small interfering RNA pathway in *Drosophila*. *Nature* **453,** 798–802 (2008).
28. Kawamura, Y. *et al. Drosophila* endogenous small RNAs bind to Argonaute 2 in somatic cells. *Nature* **453,** 793–797 (2008).
29. Okamura, K. *et al.* The *Drosophila* hairpin RNA pathway generates endogenous short interfering RNAs. *Nature* **453,** 803–806 (2008).
30. Tam, O. H. *et al.* Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* **453,** 534–538 (2008).
31. Watanabe, T. *et al.* Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* **453,** 539–543 (2008).
32. Matranga, C., Tomari, Y., Shin, C., Bartel, D. P. & Zamore, P. D. Passenger-strand cleavage facilitates assembly of siRNA into Ago2-containing RNAi enzyme complexes. *Cell* **123,** 607–620 (2005).
33. Rand, T. A., Petersen, S., Du, F. & Wang, X. Argonaute2 cleaves the anti-guide strand of siRNA during RISC activation. *Cell* **123,** 621–629 (2005).
References 32 and 33 were the first studies to demonstrate the mechanism of guide-strand selection for siRNAs.
34. Li, W. & Cha, L. Predicting siRNA efficiency. *Cell. Mol. Life Sci.* **64,** 1785–1792 (2007).
35. Tafer, H. *et al.* The impact of target site accessibility on the design of effective siRNAs. *Nature Biotechnol.* **26,** 578–583 (2008).
36. Huesken, D. *et al.* Design of a genome-wide siRNA library using an artificial neural network. *Nature Biotechnol.* **23,** 995–1001 (2005).
37. Morrissey, D. V. *et al.* Potent and persistent *in vivo* anti-HBV activity of chemically modified siRNAs. *Nature Biotechnol.* **23,** 1002–1007 (2005).
38. Robbins, M. *et al.* 2′-*O*-Methyl-modified RNAs act as TLR7 antagonists. *Mol. Ther.* **15,** 1663–1669 (2007).
39. Dowler, T. *et al.* Improvements in siRNA properties mediated by 2′-deoxy-2′-fluoro-β-D-arabinonucleic acid (FANA). *Nucleic Acids Res.* **34,** 1669–1675 (2006).
40. Watts, J. K. *et al.* 2′-Fluoro-4′-thioarabino-modified oligonucleotides: conformational switches linked to siRNA activity. *Nucleic Acids Res.* **35,** 1441–1451 (2007).
41. Fisher, M. *et al.* Inhibition of MDR1 expression with altritol-modified siRNAs. *Nucleic Acids Res.* **35,** 1064–1074 (2007).
42. Lorenz, C., Hadwiger, P., John, M., Vornlocher, H.-P. & Unverzagt, C. Steroid and lipid conjugates of siRNAs to enhance cellular uptake and gene silencing in liver cells. *Bioorg. Med. Chem. Lett.* **14,** 4975–4977 (2004).
43. Howard, K. A. *et al.* RNA interference *in vitro* and *in vivo* using a novel chitosan/siRNA nanoparticle system. *Mol. Ther.* **14,** 476–484 (2006).
44. Bitko, V., Musiyenko, A., Shulyayeva, O. & Barik, S. Inhibition of respiratory viruses by nasally administered siRNA. *Nature Med.* **11,** 50–55 (2005).
45. Li, B. J. *et al.* Using siRNA in prophylactic and therapeutic regimens against SARS coronavirus in *Rhesus macaque*. *Nature Med.* **11,** 944–951 (2005).
46. Palliser, D. *et al.* An siRNA-based microbicide protects mice from lethal herpes simplex virus 2 infection. *Nature* **439,** 89–94 (2006).
47. Rozema, D. B. *et al.* Dynamic polyconjugates for targeted *in vivo* delivery of siRNA to hepatocytes. *Proc. Natl Acad. Sci. USA* **104,** 12982–12987 (2007).
48. Bartlett, D. W., Su, H., Hildebrandt, I. J., Weber, W. A. & Davis, M. E. Impact of tumor-specific targeting on the biodistribution and efficacy of siRNA nanoparticles measured by multimodality *in vivo* imaging. *Proc. Natl Acad. Sci. USA* **104,** 15549–15554 (2007).
49. Soutschek, J. *et al.* Therapeutic silencing of an endogenous gene by systemic administration of modified siRNAs. *Nature* **432,** 173–178 (2004).
50. Zimmermann, T. S. *et al.* RNAi-mediated gene silencing in non-human primates. *Nature* **441,** 111–114 (2006).
51. Peer, D., Park, E. J., Morishita, Y., Carman, C. V. & Shimaoka, M. Systemic leukocyte-directed siRNA delivery revealing cyclin D1 as an anti-inflammatory target. *Science* **319,** 627–630 (2008).
52. Dorn, G. *et al.* siRNA relieves chronic neuropathic pain. *Nucleic Acids Res.* **32,** e49 (2004).
53. Kawasaki, Y. *et al.* Distinct roles of matrix metalloproteases in the early- and late-phase development of neuropathic pain. *Nature Med.* **14,** 331–336 (2008).
54. Dore-Savard, L. *et al.* Central delivery of Dicer-substrate siRNA: a direct application for pain research. *Mol. Ther.* **16,** 1331–1339 (2008).

55. Shishkina, G. T., Kalinina, T. S. & Dygalo, N. N. Attenuation of α2A-adrenergic receptor expression in neonatal rat brain by RNA interference or antisense oligonucleotide reduced anxiety in adulthood. *Neuroscience* **129,** 521–528 (2004).
56. Pardridge, W. M. shRNA and siRNA delivery to the brain. *Adv. Drug Deliv. Rev.* **59,** 141–152 (2007).
57. Kumar, P. *et al.* Transvascular delivery of small interfering RNA to the central nervous system. *Nature* **448,** 39–43 (2007).
This paper demonstrated the important concept that an acetylcholine-receptor-binding peptide–polyarginine conjugate can deliver siRNAs across the blood–brain barrier.
58. Song, E. *et al.* Antibody mediated *in vivo* delivery of small interfering RNAs via cell-surface receptors. *Nature Biotechnol.* **23,** 709–717 (2005).
59. McNamara, J. O. *et al.* Cell type-specific delivery of siRNAs with aptamer–siRNA chimeras. *Nature Biotechnol.* **24,** 1005–1015 (2006).
60. Chu, T. C., Twu, K. Y., Ellington, A. D. & Levy, M. Aptamer mediated siRNA delivery. *Nucleic Acids Res.* **34,** e73 (2006).
References 59 and 60 were the first to show aptamer-mediated delivery of siRNAs to a specific cellular receptor.
61. Sato, Y. *et al.* Resolution of liver cirrhosis using vitamin A-coupled liposomes to deliver siRNA against a collagen-specific chaperone. *Nature Biotechnol.* **26,** 431–442 (2008).
62. Akinc, A. *et al.* A combinatorial library of lipid-like materials for delivery of RNAi therapeutics. *Nature Biotechnol.* **26,** 561–569 (2008).
63. Brummelkamp, T. R., Bernards, R. & Agami, R. Stable suppression of tumorigenicity by virus-mediated RNA interference. *Cancer Cell* **2,** 243–247 (2002).
64. Raoul, C. *et al.* Lentiviral-mediated silencing of SOD1 through RNA interference retards disease onset and progression in a mouse model of ALS. *Nature Med.* **11,** 423–428 (2005).
65. Ralph, G. S. *et al.* Silencing mutant SOD1 using RNAi protects against neurodegeneration and extends survival in an ALS model. *Nature Med.* **11,** 429–433 (2005).
66. Carlson, M. E., Hsu, M. & Conboy, I. M. Imbalance between pSmad3 and Notch induces CDK inhibitors in old muscle stem cells. *Nature* **454,** 528–532 (2008).
67. Farah, M. H. RNAi silencing in mouse models of neurodegenerative diseases. *Curr. Drug Deliv.* **4,** 161–167 (2007).
68. Xia, H. *et al.* RNAi suppresses polyglutamine-induced neurodegeneration in a model of spinocerebellar ataxia. *Nature Med.* **10,** 816–820 (2004).
69. Shen, J. *et al.* Suppression of ocular neovascularization with siRNA targeting VEGF receptor 1. *Gene Ther.* **13,** 225–234 (2006).
70. Kleinman, M. E. *et al.* Sequence- and target-independent angiogenesis suppression by siRNA via TLR3. *Nature* **452,** 591–597 (2008).
This study found that macular vascularization could be inhibited in a non-sequence-specific manner by siRNA-mediated activation of TLR3.
71. Li, M. J. *et al.* Inhibition of HIV-1 infection by lentiviral vectors expressing Pol III-promoted anti-HIV RNAs. *Mol. Ther.* **8,** 196–206 (2003).
72. Chang, J. *et al.* Liver-specific microRNA miR-122 enhances the replication of hepatitis C virus in nonhepatic cells. *J. Virol.* **82,** 8215–8223 (2008).
73. Randall, G. *et al.* Cellular cofactors affecting hepatitis C virus infection and replication. *Proc. Natl Acad. Sci. USA* **104,** 12884–12889 (2007).
74. van Rooij, E. *et al.* Control of stress-dependent cardiac growth and gene expression by a microRNA. *Science* **316,** 575–579 (2007).
75. Calin, G. A. & Croce, C. M. MicroRNA–cancer connection: the beginning of a new tale. *Cancer Res.* **66,** 7390–7394 (2006).
76. Esau, C. C. & Monia, B. P. Therapeutic potential for microRNAs. *Adv. Drug Deliv. Rev.* **59,** 101–114 (2007).
77. Soifer, H. S., Rossi, J. J. & Saetrom, P. MicroRNAs in disease and potential therapeutic applications. *Mol. Ther.* **15,** 2070–2079 (2007).
78. Baek, D. *et al.* The impact of microRNAs on protein output. *Nature* **455,** 64–71 (2008).
79. Selbach, M. *et al.* Widespread changes in protein synthesis induced by microRNAs. *Nature* **455,** 58–63 (2008).
80. Amarzguioui, M. *et al.* Rational design and *in vitro* and *in vivo* delivery of Dicer substrate siRNA. *Nature Protoc.* **1,** 508–517 (2006).
81. Rose, S. D. *et al.* Functional polarity is introduced by Dicer processing of short substrate RNAs. *Nucleic Acids Res.* **33,** 4140–4156 (2005).
82. Kim, D. H. *et al.* Synthetic dsRNA Dicer substrates enhance RNAi potency and efficacy. *Nature Biotechnol.* **23,** 222–226 (2005).
83. Siolas, D. *et al.* Synthetic shRNAs as potent RNAi triggers. *Nature Biotechnol.* **23,** 227–231 (2005).
84. Scacheri, P. C. *et al.* Short interfering RNAs can induce unexpected and divergent changes in the levels of untargeted proteins in mammalian cells. *Proc. Natl Acad. Sci. USA* **101,** 1892–1897 (2004).
85. Jackson, A. L. *et al.* Expression profiling reveals off-target gene regulation by RNAi. *Nature Biotechnol.* **21,** 635–637 (2003).
86. Birmingham, A. *et al.* 3′ UTR seed matches, but not overall identity, are associated with RNAi off-targets. *Nature Methods* **3,** 199–204 (2006).
87. Jackson, A. L. *et al.* Position-specific chemical modification of siRNAs reduces 'off-target' transcript silencing. *RNA* **12,** 1197–1205 (2006).
88. Ui-Tei, K. *et al.* Functional dissection of siRNA sequence by systematic DNA substitution: modified siRNA with a DNA seed arm is a powerful tool for mammalian gene silencing with significantly reduced off-target effect. *Nucleic Acids Res.* **36,** 2136–2151 (2008).
89. Chiu, Y. L. & Rana, T. M. RNAi in human cells: basic structural and functional features of small interfering RNA. *Mol. Cell* **10,** 549–561 (2002).
90. Agrawal, S. & Kandimalla, E. R. Role of Toll-like receptors in antisense and siRNA. *Nature Biotechnol.* **22,** 1533–1537 (2004).
91. Judge, A. D., Bola, G., Lee, A. C. & MacLachlan, I. Design of noninflammatory synthetic siRNA mediating potent gene silencing *in vivo*. *Mol. Ther.* **13,** 494–505 (2006).
92. Schlee, M., Hornung, V. & Hartmann, G. siRNA and isRNA: two edges of one sword. *Mol. Ther.* **14,** 463–470 (2006).
93. Armstrong, M. E. *et al.* Small interfering RNAs induce macrophage migration inhibitory factor production and proliferation in breast cancer cells via a double-stranded RNA-dependent protein kinase-dependent mechanism. *J. Immunol.* **180,** 7125–7133 (2008).
94. Sioud, M. Does the understanding of immune activation by RNA predict the design of safe siRNAs? *Front. Biosci.* **13,** 4379–4392 (2008).
95. Medarova, Z., Pham, W., Farrar, C., Petkova, V. & Moore, A. *In vivo* imaging of siRNA delivery and silencing in tumors. *Nature Med.* **13,** 372–377 (2007).

Acknowledgements I thank the National Institutes of Health for grant assistance.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Correspondence should be addressed to J.J.R. (jrossi@coh.org).

ARTICLES

Changes in the phase of the annual cycle of surface temperature

A. R. Stine[1], P. Huybers[2] & I. Y. Fung[1]

The annual cycle in the Earth's surface temperature is extremely large—comparable in magnitude to the glacial–interglacial cycles over most of the planet. Trends in the phase and the amplitude of the annual cycle have been observed, but the causes and significance of these changes remain poorly understood—in part because we lack an understanding of the natural variability. Here we show that the phase of the annual cycle of surface temperature over extratropical land shifted towards earlier seasons by 1.7 days between 1954 and 2007; this change is highly anomalous with respect to earlier variations, which we interpret as being indicative of the natural range. Significant changes in the amplitude of the annual cycle are also observed between 1954 and 2007. These shifts in the annual cycles appear to be related, in part, to changes in the northern annular mode of climate variability, although the land phase shift is significantly larger than that predicted by trends in the northern annular mode alone. Few of the climate models presented by the Intergovernmental Panel on Climate Change reproduce the observed decrease in amplitude and none reproduce the shift towards earlier seasons.

Climate change is often described by trends in annual mean temperature, but large seasonal temperature changes exist independent of changes in the annual mean. A small literature exists concerning the variability in the phase of the annual cycle. Thomson[1] examined the Central England Temperature time series (1659–1990), and identified a trend in the phase of the annual cycle towards later seasons, beginning around 1950, that is anomalously large in the context of the preceding several-hundred-year record. He argued that this excursion is associated with increases in atmospheric CO_2 concentration. He also presented evidence of trends in the phase of the annual cycle over larger spatial scales and an increase in the spatial variance of the trends. Mann and Park[2] and Wallace and Osborn[3] demonstrated that the hemispheric averaged observations contain trends in amplitude and phase. The amplitude trend is negative and is related to the observation that winter is, on average, warming more quickly than summer[4–6]. The hemispheric phase trend, however, is towards earlier seasons, opposite in direction to that found by Thomson[1] for central England.

The importance of these observed amplitude and phase trends is hard to judge because we lack a good model for natural variability. Wallace and Osborn[3] used two criteria to evaluate whether the observed trends are unusual: (1) a statistical test for the presence of a trend and (2) a comparison of trends with natural variability as represented in a general circulation model. Neither of these approaches is altogether satisfactory. We expect low-frequency variability always to be present, so the presence of a trend in and of itself is not surprising[7]. Furthermore, general circulation models may not give us an accurate picture of low-frequency variability, particularly in phase, because of two shortcomings. First, the models that have been used to evaluate phase and amplitude variability have used seasonal heat and freshwater flux adjustments to match the mean annual cycle, which may artificially stabilize the modelled annual cycle. Second, and more troubling, Northern Hemisphere phase trends predicted by models forced with twentieth-century anthropogenic forcing are in the opposite direction to the observed trend[2,3]. Modelled Northern Hemisphere amplitude trends also disagree with observations when compared using a temporally fixed network[3]. Models that are unable to replicate observed trends are clearly not ideal for constraining the range of natural variability. Instead, we appeal to the early observational record to estimate the natural spatial and temporal variability of the seasonal cycle and ask if the trends observed in the recent record are anomalous in nature.

The basic state of the annual cycle

Two distinct temperature-based methods for discussing the timing of the seasons have been used in the literature. The more common is a threshold-based model wherein seasonal transitions are defined as the times of year when the temperature rises above or drops below some specific value. In this framework, the 'spring' threshold will be reached earlier if temperature increases uniformly through the year. This type of change is of first-order importance for explaining changes in seasonality observed both in biological systems (for example flowering dates[8,9], bird migration timing[8,9] and terrestrial surface carbon uptake[10]) and in components of climate that exhibit threshold responses (for example the freezing and melting of ice[11]). However, threshold-based definitions conflate changes in the phase of the annual cycle with changes in the annual mean (see Supplementary Information). We instead describe the seasonal cycle by the amplitude and phase of the yearly-period sinusoidal component in surface temperature, a measure of seasonality that is distinct from changes in the annual mean[12–16], and reference it to the yearly-period sinusoidal component in local solar insolation. The difference between the temperature and local insolation phases ($\lambda = \phi_T - \phi_{sun}$) is the lag[17], and the ratio of the amplitudes ($\mathcal{G} = A_T/A_{sun}$) is the gain (see Methods Summary). We examine gridded $5^\circ \times 5^\circ$ temperature records from the University of East Anglia's Climate Research Unit[18,19] and analyse long-term mean, detrended variability and trend fields for land (λ_{land}, $\mathcal{G}_{land}$) and ocean (λ_{ocean}, $\mathcal{G}_{ocean}$).

The spatial patterns of λ and $\mathcal{G}$ (Fig. 1a, b) are dominated by the contrast between land and ocean. The larger ocean thermal mass causes it to respond more sluggishly to oscillatory forcing than land, which results in a smaller and later oceanic response. Ocean points have a mean gain of $28\,^\circ\mathrm{C}\,(\mathrm{kW\,m^{-2}})^{-1}$ (standard deviation of point-wise

[1]Department of Earth and Planetary Science, University of California, Berkeley, California, 94720, USA. [2]Department of Earth and Planetary Sciences, Harvard University, Cambridge, Massachusetts, 02138, USA.

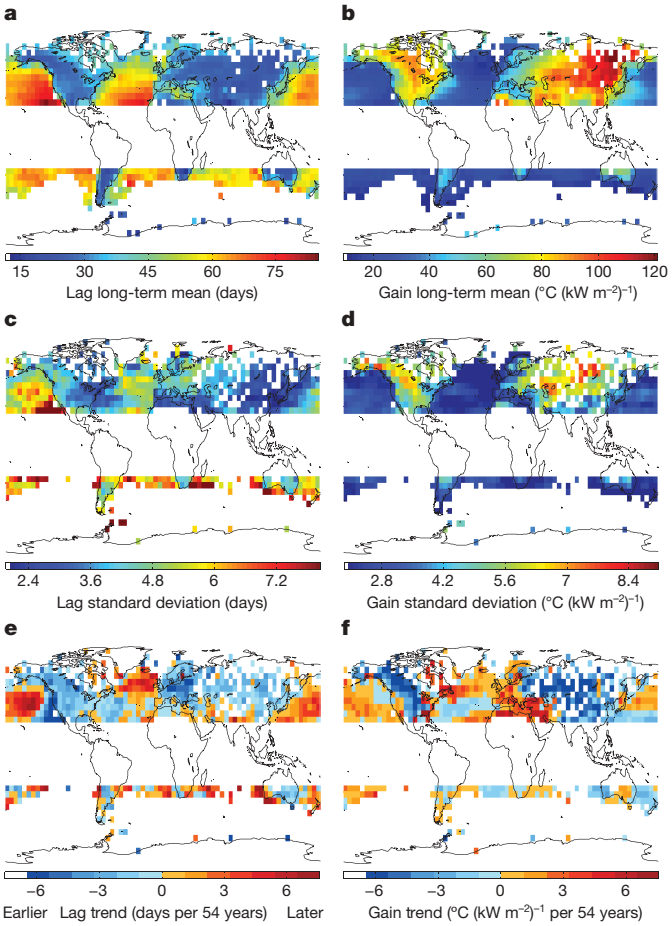


Figure 1 | Lag and gain fields. a, **c**, **e**, Phase lag, λ; **b**, **d**, **f**, amplitude gain, $\mathcal{G}$. We plot long-term-mean value (**a**, **b**), temporal standard deviation of the detrended time series (**c**, **d**); and trend in days per 54 years (**e**) and $^{\circ}\mathrm{C}\,(\mathrm{kW\,m^{-2}})^{-1}$ per 54 years (**f**). Both variability and trend maps are plotted on the 'dense network' (1954–2007), without land and ocean masks applied. Results have been excluded in the tropics, where data availability is poor, and where less than 85% of the variance in an average year is explained by the yearly component.

means, $\sigma_\mu = 15\,^{\circ}\mathrm{C}\,(\mathrm{kW\,m^{-2}})^{-1}$) and a mean lag of 56 days ($\sigma_\mu = 11$ days); in comparison, the more rapidly adjusting land has a mean gain of $74\,^{\circ}\mathrm{C}\,(\mathrm{kW\,m^{-2}})^{-1}$ ($\sigma_\mu = 23\,^{\circ}\mathrm{C}\,(\mathrm{kW\,m^{-2}})^{-1}$) and a mean lag of 29 days ($\sigma_\mu = 6$ days).

Superimposed on the dominant land–ocean contrast is an east–west gradient in $\mathcal{G}$ and λ. As we move from west to east (following the prevailing winds) across the mid-latitude continents, there is a tendency towards a more rapid response (large $\mathcal{G}_{\mathrm{land}}$, small λ_{land}). This cross-continent gradient in $\mathcal{G}_{\mathrm{land}}$ is quite strong, whereas the gradient in λ_{land} is relatively weak (λ_{land} adjusts rapidly to interior values along the western continental margin[1,2]). Conversely, as we move from west to east across the mid-latitude ocean basins, there is a tendency towards a more sluggish response (small $\mathcal{G}_{\mathrm{ocean}}$, large λ_{ocean}), and the relative strengths of the $\mathcal{G}_{\mathrm{ocean}}$ and λ_{ocean} gradients is reversed relative to that of the land ($\mathcal{G}_{\mathrm{ocean}}$ adjusts rapidly to interior values along the western margin of ocean basins).

The role of land–sea contrast in setting the climatological distribution of the annual cycle is not a new observation[17,20–22], but its dominance is particularly obvious when considering the relationship between $\mathcal{G}$ and λ. Pairs of $\mathcal{G}$ and λ fall along an arc (Fig. 2a). We define a 'seasonal response index' to represent a point's position in this lag–gain space as

$$\mathrm{SRI} = \frac{\mathcal{G} - \min(\mathcal{G})}{\max(\mathcal{G}) - \min(\mathcal{G})} - \frac{\lambda - \min(\lambda)}{\max(\lambda) - \min(\lambda)}$$

and find that 75% of the variance in this index is explained by the distance between a grid point and the coast to its west, where distance is taken as positive for land and negative for ocean (see Supplementary Information for more discussion on the structure of variability). The relationship between SRI and distance from the coast holds best in Eurasia, where southern mountains constrain the transport to be zonal and isolate the interior from tropical moisture. Deviations from this general east–west pattern are found in regions where transport is less zonal, such as the southeastern North American monsoonal region, where there is strong poleward moisture transport onto land, and in the western United States, where the north–south alignment of the Rocky Mountains effectively blocks oceanic influence from the Pacific Ocean.

The observed arc in the relationship between $\mathcal{G}$ and λ is a ubiquitous feature of seasonally driven models that contain interacting land and ocean regions, and can be understood as the natural consequence of interacting sinusoids. Consider two sine waves with different phases and amplitudes, $S_1 = A\sin(\phi)$ and $S_2 = (A/r)\sin(\phi + \Delta\phi)$. A weighted average of the two sine waves, $wS_1 + (1-w)S_2$, with $0 < w < 1$, yields a sine wave with amplitude

$$A_w = \frac{A}{r}\sqrt{w^2r^2 + 2rw(1-w)\cos\Delta\phi + (1-w)^2}$$

and phase

$$\phi_w = \phi + \cos^{-1}\left(\frac{A}{A_w}\left(w - \frac{1}{r}(1-w)\cos\Delta\phi\right)\right)$$

A mixing line using this equation (Fig. 2a) is consistent with the general form of the observed arc. Apparently, the spatial structure associated with the seasonal cycle can be understood, to first order, as the result of variable mixing between continental and marine influence.

Variability in $\mathcal{G}$ (Fig. 1d) tends to be largest where the climatological $\mathcal{G}$ is large (coefficient of correlation, $R = 0.83$), and is about twice as large over land (mean of point-wise standard deviations, $\bar{\sigma} = 5.2\,^{\circ}\mathrm{C}\,(\mathrm{kW\,m^{-2}})^{-1}$) as it is over the ocean ($\bar{\sigma} = 2.5\,^{\circ}\mathrm{C}\,(\mathrm{kW\,m^{-2}})^{-1}$). Conversely, temporal variability in λ (Fig. 1c) is correlated with $\mathcal{G}^{-1}$ ($R = 0.62$) and is larger over the ocean ($\bar{\sigma} = 5.0$ days) than it is over land ($\bar{\sigma} = 4.0$ days). The inverse relationship and larger λ_{ocean} variability arises because finite perturbations more readily alter the phase of a smaller amplitude sinusoid (see the Supplementary Information discussion on natural variability). We thus expect that it will be more difficult to detect the presence of any true phase trend over the ocean.

Trends in the phase and gain of the annual cycle

The 1954–2007 λ_{land} trends (Fig. 1e) are predominantly towards earlier seasons, with a mean decrease of 1.7 days (that is, 6%) over the past 54 years. The λ_{ocean} trends are large but regionally disparate. For example, the interior of the North Pacific, and the Atlantic north of 50° N, exhibit trends towards later seasons, whereas along the eastern edge of the North Pacific, and in the North Atlantic south of 50° N, trends are primarily towards earlier seasons. The mean λ_{ocean} shift is towards later seasons by 1.0 days over the past 54 years.

A comparison of trend maps (Fig. 1e, f) and variability maps (Fig. 1c, d) reveals that the trends are large where the detrended variability is large. This suggests the obvious null hypothesis that the trends are merely a manifestation of natural variability. One test for whether the trends observed in the recent record are consistent with natural variability is to compare them with trends observed in earlier periods. We consider the distribution of point-wise trends (Fig. 2b) for the 1900–1953 and 1954–2007 intervals using those records which have good temporal coverage during both intervals (see Methods; this is the default distribution of records that we use below, unless specifically stated otherwise). Land and ocean are considered separately because the characters of their annual cycles are so different.

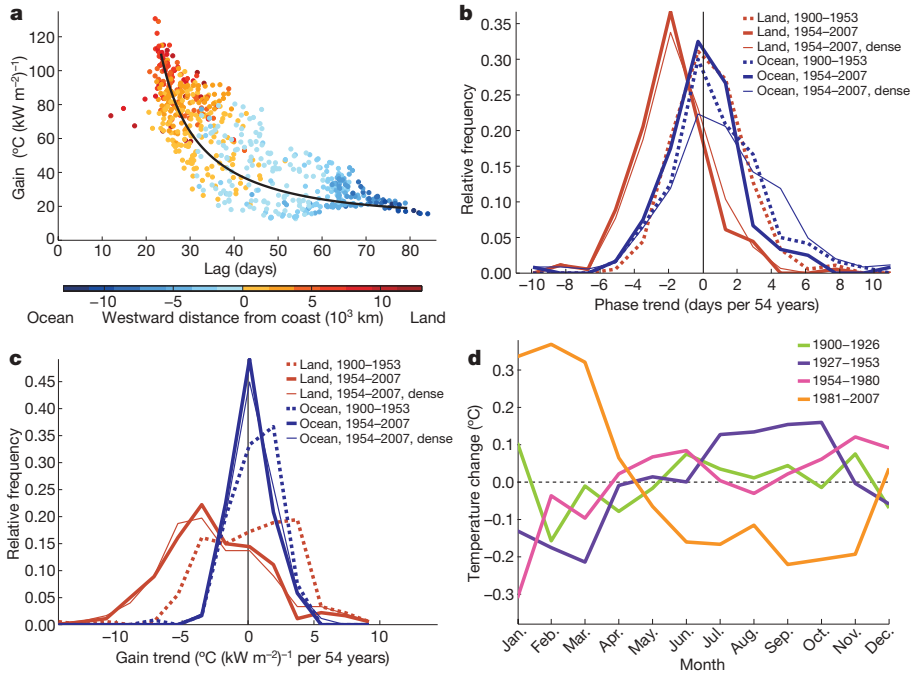


Figure 2 | Mean annual cycle and distribution and character of trends. **a**, Observed relationship between local gain, $\mathcal{G}$, and lag, λ, for Northern Hemisphere extratropical locations. Colour represents the distance between a grid point and the coast to its west (positive for land, negative for ocean). Outliers with $\lambda < 20$ days are from the Indian subcontinent and presumably reflect monsoon dynamics. The black line shows the nonlinear relationship between amplitude and phase for weighted averages of two end-member sine waves (see Supplementary Information). **b**, Normalized histograms of point-wise λ trends (in days per 54 years). Red lines represent land; blue lines represent ocean. The dotted lines give the distributions for the control period (1900–1953) on the 'comparison network'. The thick solid lines give the distributions for the same spatial network for 1954–2007. The thin solid line is for the dense network (1954–2007; see Methods for network descriptions). **c**, Same as **b**, but for $\mathcal{G}$ trends (in $^{\circ}\mathrm{C}\,(\mathrm{kW\,m^{-2}})^{-1}$ per 54 years). **d**, Anomalies in composite land annual-cycle shape for 27-year periods, relative to the 108-year composite. Southern Hemisphere grid boxes have been shifted by six months before averaging.

We adopt a null hypothesis that the mean of each distribution of trends is zero. Testing this null hypothesis requires a knowledge of the effective spatial degrees of freedom[23,24], and we use estimates obtained from the moment-matching method of ref. 25 (see Methods). Of the four distributions of λ trends that we consider, only those over land during the 1954–2007 interval have a mean that differs significantly from zero (Table 1), and here the significance is marked (-1.9 ± 1.0 days per 54 years, $P < 0.001$). Repeating the tests for 1954–2007, using the larger spatial network that is available for this interval (the dense network; see Methods), supports the significance of the λ_{land} trend (-1.7 ± 0.8 days per 54 years, $P < 0.001$). We also detect a significant 1954–2007 λ_{ocean} trend towards later seasons (1.0 ± 0.9 days per 54 years, $P = 0.02$) that is only detectable in the more extensive dense network.

The dominant signal in the 1954–2007 $\mathcal{G}$ trend (Fig. 1f) is a decrease in the amplitude of the annual cycle over land, averaging $-2.5\,^{\circ}\mathrm{C}\,(\mathrm{kW\,m^{-2}})^{-1}$ over the past 54 years on the dense network, a 3% drop. This is the well-known amplification of winter warming[4,5,26], which is strongest in the interior of Eurasia and in the boreal forests of western Canada. Note, however, that large regions exist where the amplitude has increased. In western Europe and the Middle East, the observed increase in $\mathcal{G}$ is associated with greater warming in summer than in winter. In the central North Pacific and the south-eastern United States, the increase in $\mathcal{G}$ results from winter cooling. In Quebec, the summer warming and winter cooling trends are of comparable magnitudes, leaving little trend in mean temperature but a measurable increase in $\mathcal{G}$. Ocean $\mathcal{G}$ trends are almost everywhere small and show a mean increase of $0.4\,^{\circ}\mathrm{C}\,(\mathrm{kW\,m^{-2}})^{-1}$ over the past 54 years.

We apply tests to the $\mathcal{G}$ trends similar to those made on the λ trends (Fig. 2c). Of the four distributions of $\mathcal{G}$ trends, only those over land during 1954–2007 have a mean that differs significantly from zero ($-2.6 \pm 2.4\,^{\circ}\mathrm{C}\,(\mathrm{kW\,m^{-2}})^{-1}$ per 54 years, $P < 0.03$; Table 1). If the dense network is used instead for this interval, the $\mathcal{G}_{\mathrm{land}}$ trends remain significant ($P = 0.05$) and the $\mathcal{G}_{\mathrm{ocean}}$ trends emerge as being marginally significant ($P = 0.07$). It is noteworthy that the well-reported changes

Table 1 | Means of trend distributions

	Comparison network		Dense network
	1900–1953	1954–2007	1954–2007
λ_{land}	0.31 ± 0.95 ($P = 0.5$)	**-1.88 ± 0.95 ($P < 0.001$)**	**-1.66 ± 0.81 ($P < 0.001$)**
λ_{ocean}	0.72 ± 1.24 ($P = 0.2$)	0.00 ± 1.23 ($P = 1$)	**1.02 ± 0.87 ($P = 0.02$)**
$\mathcal{G}_{\mathrm{land}}$	0.23 ± 2.21 ($P = 0.8$)	**-2.62 ± 2.40 ($P = 0.03$)**	**-2.54 ± 2.54 ($P = 0.05$)**
$\mathcal{G}_{\mathrm{ocean}}$	0.44 ± 0.82 ($P = 0.3$)	0.27 ± 0.69 ($P = 0.4$)	0.43 ± 0.47 ($P = 0.07$)
Summer land temp.	**0.54 ± 0.37 ($P = 0.007$)**	**0.86 ± 0.38 ($P < 0.001$)**	**0.96 ± 0.35 ($P < 0.001$)**
Summer ocean temp.	**0.79 ± 0.28 ($P < 0.001$)**	**0.60 ± 0.26 ($P < 0.001$)**	**0.48 ± 0.18 ($P < 0.001$)**
Winter land temp.	0.60 ± 0.96 ($P = 0.2$)	**1.66 ± 1.00 ($P = 0.005$)**	**1.77 ± 1.40 ($P = 0.02$)**
Winter ocean temp.	**0.72 ± 0.33 ($P < 0.001$)**	**0.46 ± 0.29 ($P = 0.005$)**	**0.41 ± 0.19 ($P < 0.001$)**
Summer–winter land temp.	-0.06 ± 0.83 ($P = 0.9$)	-0.80 ± 0.95 ($P = 0.09$)	-0.81 ± 0.94 ($P = 0.08$)
Summer–winter ocean temp.	0.07 ± 0.32 ($P = 0.6$)	0.14 ± 0.28 ($P = 0.3$)	0.07 ± 0.21 ($P = 0.5$)

Phase lag (λ) trends are in expressed in days per 54 years, amplitude gain ($\mathcal{G}$) trends are expressed in $^{\circ}\mathrm{C}\,(\mathrm{kW\,m^{-2}})^{-1}$ per 54 years and temperature (temp.) trends are expressed in °C per 54 years. Two-tailed P values are given in parentheses. Significant values, judged using 95% confidence intervals, are set in bold.

in the amplitude of the annual cycle[19,27] are less significant than the less-reported land-phase trend. The low significance of the amplitude results is related to the large natural variability in wintertime temperature. Winter warming is considerably stronger than summer warming over land during 1954–2007, but the variance in winter land temperatures is about four times that in summer land temperatures, making the winter trend less significant and making detection of changes in amplitude difficult[28]. In fact, we are unable to detect a significant difference between summer and winter warming when temperature trends are analysed as the difference between three-month seasonal averages (Table 1).

We focus on the λ_{land} trends because their significance is markedly higher than that of any other observed trend. Furthermore, $\mathcal{G}$ trends are more easily discussed in the seasonal-average-temperature framework than are λ trends and have received more attention elsewhere[5,6,27].

There are two steps in establishing the presence of an anomalous trend. The first is establishing that a trend is statistically distinguishable from zero, which we demonstrated for the 1954–2007 λ_{land} observations. The second is establishing that this trend is different in character from what we would expect in the naturally varying system, which is more difficult given the finite length of the instrumental temperature records. We are particularly concerned about low-frequency variability being misinterpreted as an anomalous trend[7]. The absence of a significant λ_{land} trend for the 1900–1953 test period indicates that the trend during 1954–2007 is anomalous. By restricting ourselves to a smaller set of locations, we can also extend our analysis back to 1850. We construct an average λ_{land} time series by averaging the phase time series from all of the land grid boxes with perfect temporal coverage between 1850 and 2007, and adopt the null hypothesis that the 1954–2007 trends result from natural low-frequency variability as represented in the 1850–1953 record. We build a distribution for this null hypothesis by calculating the trends of many synthetic time series having the same spectral amplitude structure as the 1850–1953 record, but with randomized phases[29] (see Supplementary Methods), and are able to reject it with very high confidence ($P = 0.006$). The phase trend over the past 54 years is not consistent with the structure of natural variability found in the earlier record. Furthermore, there is no 54-year period in the 1850–1953 control period that would allow rejection of this null hypothesis. (Note that we are unable to meaningfully compare the 1954–2007 trends in λ_{ocean} with the 1850–1953 period because instrumental coverage over the ocean during these early times was poor.)

Finally, it is reasonable to ask whether the observed variability is really best thought of as a shift in the yearly sinusoidal component, or if it would be better described by changes in individual months. A change in a single month's temperature will map into a shift in the annual cycle, although the yearly frequency component provides a poor description of such an anomaly. We calculate the mean annual cycles for four 27-year periods, using land grid boxes with good temporal coverage between 1900 and 2007 (see Supplementary Methods), and consider their anomalies from the 108-year mean annual cycle (Fig. 2d). (Consideration of the means in these four periods gives insight into the origins of the trends in the 1900–1953 and 1954–2007 intervals.) The most recent anomaly time series (1981–2007) exhibits the largest departures from the long-term mean, and 80% of its variance is explained by the yearly component. The most recent period has more variability at the annual period than the total variability during all preceding periods, highlighting both that these shifts are well described by a yearly sinusoidal component and the anomalous nature of the recent changes (see Supplementary Table 2).

Origins of the changes in the annual cycle

To explore the origins of the shifts in $\mathcal{G}_{land}$ and λ_{land}, we first turn to the global climate model results summarized in the Fourth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC)[30]. In particular, we analyse the 72 simulations of twentieth-century climate that use observed forcings conducted for the World Climate Research Programme's (WCRP) Coupled Model Intercomparison Project Phase 3 (ref. 31). The distributions of mean trends in $\mathcal{G}_{land}$ and λ_{land} found in the models easily encompass the observed trends during the 1900–1953 interval. However, during the 1954–2000 interval (the simulations stop at the end of the century), the observed decrease in $\mathcal{G}_{land}$ has a larger magnitude than all but six of the model simulations, and no model reproduces the observed shift towards earlier seasons (Fig. 3). The mean of the model λ_{land} trends for 1954–2000 is towards later, not earlier, seasons. Furthermore, 25 atmospheric model runs forced with observed sea surface temperatures for the 1978–2000 period (the Atmospheric Model Intercomparison Project models[32]) do no better at replicating the observed λ_{land} trends (see Supplementary Methods and Discussion).

The IPCC model results do not appear to give us an explanation of the observed trends, except to suggest that the answer involves something that the models do not capture. We thus retreat to a simple, conceptual model to explore how local processes may cause variability in λ and $\mathcal{G}$ consistent with the observations. Our goal here is to explore some obvious candidates and roughly estimate the size of the perturbation needed to explain the observation. We use a one-box energy

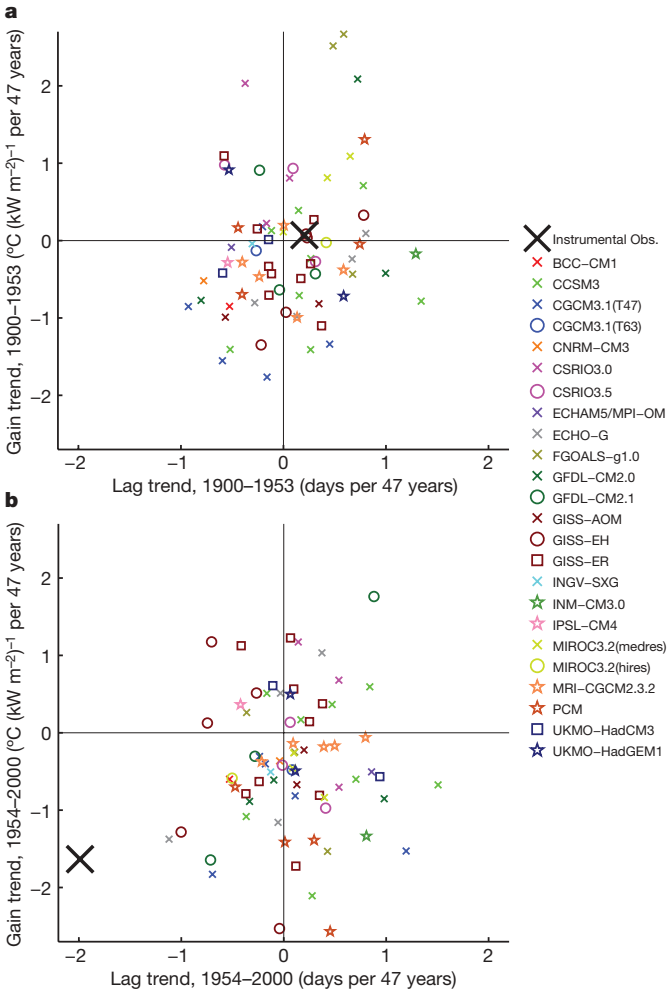


Figure 3 | Modelled and observed mean land trends. **a**, Observed 1900–1953 land lag, λ_{land}, and gain, $\mathcal{G}_{land}$, trends and those from WCRP 'Climate of the Twentieth Century' simulations, sampled at the same locations. Marks with same colour and shape indicate multiple runs with the same model. The large black cross indicates the actual observed trends ('Instrumental Obs.'). **b**, Same as **a**, but for 1954–2000. The comparison ends at 2000 because IPCC runs generally stop then. Individual model descriptions are given in ref. 50 and references therein.

balance model based on ref. 33, forced with sinusoidally varying short-wave radiation, with atmospheric short-wave optical properties calculated following ref. 34 (see Methods).

Many of the mechanisms invoked to explain variability in annual mean temperature are unlikely to be directly responsible for the observed shift in phase. Doubling the atmospheric long-wave optical depth to simulate the radiative effect of a very large increase in greenhouse gases has essentially no effect on seasonal timing ($\Delta\lambda = 0.1$ days, $\Delta\mathcal{G} = -0.6\,^{\circ}\mathrm{C}\,(\mathrm{kW\,m^{-2}})^{-1}$). Increasing solar luminosity by a fixed percentage increases the amplitude of the temperature response by the same percentage and has a negligible effect on phase[1,35]. Decreases in sea ice (not represented in our model) present the atmosphere with a larger thermal mass, implying a delayed seasonal response (although threshold responses at the time of spring melt may induce changes in the opposite direction). Consistent with this intuition, the (incorrect) phase delays found in the model results of ref. 2 are attributed to decreases in sea-ice cover.

However, there exist numerous mechanisms that may shift the seasonal cycle in the observed direction. A decrease in thermal mass on land of $8 \pm 4\%$ is sufficient to produce the observed offset in phase of 1.7 ± 0.8 days. Thermal mass on land is largely modulated by soil moisture. Compare, for example, the effective thermal mass of a dry desert sand ($1.9\,\mathrm{J\,m^{-3}\,^{\circ}C^{-1}}$) with that of a saturated loam soil ($3.2\,\mathrm{J\,m^{-3}\,^{\circ}C^{-1}}$). For a typical soil, the observed phase shift would require a $13 \pm 7\%$ decrease in soil moisture. IPCC Fourth Assessment Report model runs disagree with each other on the sign of recent soil moisture trends and show little skill at explaining the (sparse) observations[36]. The few available long-term measurements suggest increased soil moisture over the latter part of the twentieth century[37,38], which is inconsistent with the thermal mass hypothesis, although we observe that drought reconstructions[39] indicate these observations may not be representative of continental-scale variations. The paucity of records with more than 40 years of data prohibits a more detailed comparison. We consider large-scale decreases in soil moisture to be a viable candidate for inducing the observed shift towards earlier seasons.

Perturbations to atmospheric short-wave optical properties are also effective at modifying the annual cycle, and it appears that short-wave absorptivity has been changing, perhaps because of aerosols[40–42]. The Earth's short-wave optical properties are not constant throughout the year, and their potential range of variability is not captured by this simple model. Nonetheless, the model indicates that variability in atmospheric annual mean reflectivity, absorptivity or transmissivity on the order of 10% will change λ_{land} by the observed amount. Note that Wallace and Osborn[3] were unable to replicate the observed hemispheric phase shifts using a general circulation model, but that the inclusion of aerosol forcing did decrease the modelled (incorrect) shift towards later seasons. We see no indication of shifts in mean λ_{land} after any of the major volcanic eruptions of the past century, although the effects of stratospheric and tropospheric aerosols on phase are likely to be quite different.

Thomson[1] makes the intriguing, although difficult-to-evaluate, proposal that decreases in phase are due to an increased local sensitivity to anomalistic year forcing (associated with the annual cycle in Earth–Sun distance) relative to tropical year forcing (due to the annual cycle in the orientation of the Earth's rotation axis relative to the Sun).

The above-mentioned changes in albedo, soil moisture and short-wave forcing have all been implicated in changing modes of atmospheric circulation[43–45]. This raises the further possibility that shifts in atmospheric circulation participate in the modulation of the annual cycle. We focus on the northern annular mode (NAM) and the Pacific North American pattern, as these have been shown to represent the bulk of the variability in standard atmospheric climate indices[46] (but see Supplementary Information for a more complete analysis). The NAM shows significant cross-correlations with time series of 1950–2007 spatially averaged λ_{land} ($R = -0.5$, $P < 0.001$) and $\mathcal{G}_{\mathrm{land}}$ ($R = 0.42$, $P = 0.007$), whereas the Pacific North American pattern has significant correlation with $\mathcal{G}_{\mathrm{ocean}}$ ($R = 0.3$, $P = 0.04$). Apparently, atmospheric dynamical processes respond to similar forcing mechanisms as λ_{land} or themselves participate in altering λ_{land} through the advection of heat and moisture or other indirect processes. Northern Hemisphere snow cover, for example, is known to interact with the NAM[43,47], and wind-driven changes in mixed layer depth affect the thermal mass that the ocean presents to the atmosphere[48].

The Atmospheric Model Intercomparison Project simulations have been found to capture the spatial pattern, but not the temporal pattern, of NAM variability[49], just as we find that the models fail to capture the long-term trends in phase. However, the recent phase excursion appears to be only partly explained by the late-twentieth-century excursion in the NAM. A regression of the spatial average λ_{land} time series against the NAM index from 1950–2007 removes 25% of the variance and 40% of the trend in the λ_{land} time series, but still leaves a significant trend ($P < 0.02$) of -1.0 days per 57 years (see Supplementary Information).

The statistics of the distribution of λ_{land} trends are well described as natural variability from 1900–1953, but the distribution shifts in 1954–2007, the period in which anthropogenic interference with mean temperature becomes apparent. If we extend our natural control period back to 1850, the recent trends appear yet more anomalous. Numerous climate factors can influence the phase of the annual cycle, and it appears that some portion of the trend in the annual cycle is associated with changes in the NAM during the late twentieth century. We expect that a complete explanation for long-term shifts in atmospheric circulation will also encompass an explanation of the variability in the phase of the annual cycle. Although the mechanism is still uncertain, the tests we apply to the 1954–2007 trends in land phase indicate that they are inconsistent with natural variability, and thus appear to be due to anthropogenic influence.

METHODS SUMMARY

For each year of data, we calculate the yearly (one cycle per year) sinusoidal component using the Fourier transform, as

$$Y_x = \frac{2}{12}\sum_{t=0.5}^{11.5} \mathrm{e}^{2\pi i t/12} x(t+t_0)$$

where $x(t+t_0)$, $t = 0.5, \ldots, 11.5$, are 12 monthly values of either the de-meaned monthly temperature or de-meaned monthly insolation and t is time in months. The factor of two is to account for both positive and negative frequencies. Phase is given by $\phi_x = \tan^{-1}(\mathrm{Im}(Y_x)/\mathrm{Re}(Y_x))$. To discuss both hemispheres in a common framework, we reference the temperature phase, ϕ_{T}, to the local solar insolation phase, ϕ_{sun}. The difference between these two phases is the lag[17], $\lambda = \phi_{\mathrm{T}} - \phi_{\mathrm{sun}}$.

Amplitude is given by $A_x = |Y_x|$. For the purpose of understanding the response of the Earth's temperature to forcing, we examine the gain, which is defined as the ratio of the amplitudes of the annual cycles in temperature and insolation, $\mathcal{G} = A_{\mathrm{T}}/A_{\mathrm{sun}}$. Unlike insolation or temperature amplitudes alone, $\mathcal{G}$ has very little latitude dependence.

If any of the 12 monthly temperature values is missing in the data set at a given location, then no estimate of the annual cycle is made at that location for that year. Analyses using longer record pieces and more advanced filter techniques do not change our conclusions regarding the significance of phase and amplitude changes.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 25 April; accepted 20 November 2008.

1. Thomson, D. The seasons, global temperature, and precession. *Science* **268,** 59–68 (1995).
2. Mann, M. & Park, J. Greenhouse warming and changes in the seasonal cycle of temperature: Model versus observations. *Geophys. Res. Lett.* **23,** 1111–1114 (1996).
3. Wallace, C. & Osborn, T. Recent and future modulation of the annual cycle. *Clim. Res.* **22,** 1–11 (2002).
4. Wallace, J., Zhang, Y. & Renwick, J. Dynamic contribution to hemispheric mean temperature trends. *Science* **270,** 780–783 (1995).
5. Balling, R., Michaels, P. & Knappenberger, P. Analysis of winter and summer warming rates in gridded temperature time series. *Clim. Res.* **9,** 175–181 (1998).

6. Trenberth, K. *et al.* in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Solomon, S. D. *et al.*) 235–336 (Cambridge Univ. Press, 2007).
7. Wunsch, C. The interpretation of short climate records, with comments on the North Atlantic and Southern Oscillations. *Bull. Am. Meteorol. Soc.* **80,** 245–255 (1999).
8. Schwartz, M., Ahas, R. & Aasa, A. Onset of spring starting earlier across the Northern Hemisphere. *Glob. Change Biol.* **12,** 343–351 (2006).
9. Sparks, T. & Menzel, A. Observed changes in seasons: An overview. *Int. J. Climatol.* **22,** 1715–1725 (2002).
10. Myneni, R., Keeling, C., Tucker, C., Asrar, G. & Nemani, R. Increased plant growth in the northern high latitudes from 1981 to 1991. *Nature* **386,** 698–702 (1997).
11. Magnuson, J. *et al.* Historical trends in lake and river ice cover in the Northern Hemisphere. *Science* **289,** 1743–1746 (2000).
12. White, G. & Wallace, J. Global distribution of annual and semiannual cycles in surface-temperature. *Mon. Weath. Rev.* **106,** 901–906 (1978).
13. Thompson, R. Complex demodulation and the estimation of the changing continentality of Europe climate. *Int. J. Climatol.* **15,** 175–185 (1994).
14. Hsu, C. & Wallace, J. Global distribution of annual and semiannual cycles in precipitation. *Mon. Weath. Rev.* **104,** 1093–1101 (1976).
15. van Loon, H. in *Meteorology of the Southern Hemisphere* (ed. Newton, C. W.) 25–58 (Meteorological Monographs 35, Am. Meteorol. Soc., 1972).
16. Eliseev, A., Mokhov, I. & Guseva, M. Sensitivity of amplitude–phase characteristics of the surface air temperature annual cycle to variations in annual mean temperature. *Izvestiya. Atmos. Ocean. Phys.* **42,** 300–312 (2006).
17. Prescott, J. & Collins, J. The lag of temperature behind solar radiation. *Q. J. R. Meteorol. Soc.* **77,** 121–126 (1951).
18. Brohan, P., Kennedy, J., Harris, I., Tett, S. & Jones, P. Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850. *J. Geophys. Res.* **111,** doi:10.1029/2005JD006548 (2006).
19. Jones, P. D., New, M., Parker, D. E., Martin, S. & Rigor, I. G. Surface air temperature and its changes over the past 150 years. *Rev. Geophys.* **37,** 173–199 (1999).
20. Ward, R. The classification of climates: I. *Bull. Am. Geogr. Soc.* **38,** 401–412 (1906).
21. Kendrew, W. *Climate of the Continents* 5th edn (Oxford Univ. Press, 1961).
22. Jain, S., Lall, U. & Mann, M. Seasonality and interannual variations of Northern Hemisphere temperature: Equator-to-pole gradient and ocean-land contrast. *J. Clim.* **12,** 1086–1100 (1999).
23. Jones, P., Osborn, T. & Briffa, K. Estimating sampling errors in large-scale temperature averages. *J. Clim.* **10,** 2548–2568 (1997).
24. Madden, R., Shea, D., Branstator, G., Tribbia, J. & Weber, R. The effects of imperfect spatial and temporal sampling on estimates of the global mean temperature - experiments with model data. *J. Clim.* **6,** 1057–1066 (1993).
25. Bretherton, C., Widmann, M., Dymnikov, V., Wallace, J. & Blade, I. The effective number of spatial degrees of freedom of a time-varying field. *J. Clim.* **12,** 1990–2009 (1999).
26. Parker, D., Jones, P., Folland, C. & Bevan, A. Interdecadal changes of surface-temperature since the late-19th-century. *J. Geophys. Res.* **99,** 14373–14399 (1994).
27. Wallace, J., Zhang, Y. & Bajuk, L. Interpretation of interdecadal trends in Northern Hemisphere surface air temperature. *J. Clim.* **9,** 249–259 (1996).
28. Wigley, T. M. L. & Jones, P. D. Detecting CO_2-induced climate change. *Nature* **292,** 205–208 (1991).
29. Schreiber, T. & Schmitz, A. Surrogate time series. *Physica D* **142,** 346–382 (2000).
30. Meehl, G. A. *et al.* in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Solomon, S. D. *et al.*) 747–845 (Cambridge Univ. Press, 2007).
31. Meehl, G. A. *et al.* The WCRP CMIP3 multimodel dataset - A new era in climate change research. *Bull. Am. Meteorol. Soc.* **88,** 1383–1394 (2007).
32. Gates, L. AMIP: The Atmospheric Model Intercomparison Project. *Bull. Am. Meteorol. Soc.* **73,** 1962–1970 (1992).
33. Goody, R. M. *Principles of Atmospheric Physics and Chemistry* Ch. 5 (Oxford Univ. Press, 1995).
34. Shell, K. & Somerville, R. A generalized energy balance climate model with parameterized dynamics and diabatic heating. *J. Clim.* **18,** 1753–1772 (2005).
35. Karl, T., Jones, P. & Knight, R. Testing for bias in the climate record. *Science* **271,** 1879–1880 (1996).
36. Li, H., Robock, A. & Wild, M. Evaluation of Intergovernmental Panel on Climate Change Fourth Assessment soil moisture simulations for the second half of the twentieth century. *J. Geophys. Res.* **112,** doi:10.1029/2006JD007455 (2007).
37. Robock, A. *et al.* The global soil moisture data bank. *Bull. Am. Meteorol. Soc.* **81,** 1281–1299 (2000).
38. Vinnikov, K. & Yeserkepova, I. Soil-moisture - empirical-data and model results. *J. Clim.* **4,** 66–79 (1991).
39. Dai, A., Trenberth, K. & Qian, T. A global dataset of Palmer Drought Severity Index for 1870–2002: Relationship with soil moisture and effects of surface warming. *J. Hydrometeorol.* **5,** 1117–1130 (2004).
40. Wild, M. *et al.* From dimming to brightening: Decadal changes in solar radiation at Earth's surface. *Science* **308,** 847–850 (2005).
41. Liepert, B. Observed reductions of surface solar radiation at sites in the United States and worldwide from 1961 to 1990. *Geophys. Res. Lett.* **29,** doi:10.1029/2002GL014910 (2002).
42. Stanhill, G., Trenberth, K. & Cohen, S. A review of the evidence for a widespread and significant reduction in global radiation with discussion of its probable causes and possible agricultural consequences. *Agric. For. Meteorol.* **107,** 255–278 (2001).
43. Cohen, J. & Entekhabi, D. Eurasian snow cover variability and Northern Hemisphere climate predictability. *Geophys. Res. Lett.* **26,** 345–348 (1999).
44. Yeh, T.-C., Wetherald, R. T. & Manabe, S. The effect of soil moisture on the short-term climate and hydrology change—a numerical experiment. *Mon. Weath. Rev.* **112,** 474–490 (1984).
45. Lohmann, U. & Feichter, J. Global indirect aerosol effects: a review. *Atmos. Chem. Phys.* **5,** 715–737 (2005).
46. Quadrelli, R. & Wallace, J. A simplified linear framework for interpreting patterns of Northern Hemisphere wintertime climate variability. *J. Clim.* **17,** 3728–3744 (2004).
47. Saito, K. & Cohen, J. The potential role of snow cover in forcing interannual variability of the major Northern Hemisphere mode. *Geophys. Res. Lett.* **30,** doi:10.1029/2002GL016341 (2003).
48. Kara, A., Rochford, P. & Hurlburt, H. Mixed layer depth variability over the global ocean. *J. Geophys. Res.* **108,** doi:10.1029/2000JC000736 (2003).
49. Cohen, J., Frei, A. & Rosen, R. The role of boundary conditions in AMIP-2 simulations of the NAO. *J. Clim.* **18,** 973–981 (2005).
50. Randall, D. *et al.* in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Solomon, S. D. *et al.*) 589–662 (Cambridge Univ. Press, 2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Tingley, A. Swann and L. Morgan for comments that improved the manuscript. A.R.S. was funded in part by a Chancellor's Fellowship from the University of California. P.H. was funded in part by US National Science Foundation award 0645936. I.Y.F. acknowledges support from US National Science Foundation award 0628278. We acknowledge the Program for Climate Model Diagnosis and Intercomparison and the WCRP's Working Group on Coupled Modelling for their roles in making available the WCRP Coupled Model Intercomparison Project Phase 3 multi-model simulations. Support for these simulations is provided by the Office of Science, US Department of Energy.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to A.R.S. (zan@atmos.berkeley.edu).

METHODS

Data sets. When plotting full spatial fields (Figs 1, 2a), we use the HadCRUT3 blended land-and-ocean $5^\circ \times 5^\circ$ gridded surface temperature anomalies[18] plus gridded climatology[19] from the Climate Research Unit at the University of East Anglia. For comparison with IPCC model archive output (Fig. 3), we use the HadCRUT3 data set (because models do not calculate separate land and ocean surface temperatures), but restrict ourselves to grid points that are more than 50% land. When land and ocean are considered separately (all other calculations), we instead use the CRUTEM3 (for land) and HadSST2 (for ocean) data sets so that the land and ocean signals are more cleanly separated. Dominantly land and ocean boxes are identified using the Clark US Navy Fleet Numerical Oceanographic Center Land/Ocean Mask[51]. Because a grid box with land cover of only a few per cent is not as representative of land as a continental interior box, only CRU grid points that are more than 50% land in the Navy mask are used in land calculations. Those that are less than 50% land are considered ocean.

Data mask. We desire a high ratio of signal (annual amplitude) to noise (errors in observations and high frequency temperature variability), so that we can isolate variability that is associated with changes in the annual cycle. In fact, the yearly sinusoidal component dominates the extratropical records; on average, it explains 96% and 90% of the within-year variance in monthly temperatures for land and ocean grid boxes, respectively. With the time series of yearly $\mathcal{G}$ and λ at each grid box, we estimate the long-term means, the long-term trends and the standard deviations of the departures from the long-term trend. To do so, we exclude from analysis (1) those extratropical grid boxes where less than 85% of the average within-year variance is explained by the yearly sinusoidal component (primarily the Southern Ocean) and (2) all tropical grid boxes (23.5° S to 23.5° N), because the two-cycles-per-year harmonic in forcing and response is strong in this region. For calculating long-term-mean λ and $\mathcal{G}$, we exclude grid boxes with fewer than ten yearly estimates over the entire record. For calculating 54-year trends and detrended standard deviation in $\mathcal{G}$ and λ, we exclude grid boxes with fewer than 40 yearly estimates. Trends are calculated using a least-squares fit. For comparing 1900–1953 and 1954–2007 trends, we use a 'comparison network' of grid boxes that meet these data-density criteria for both periods (180 land points, 120 ocean grid points). We also use a 'dense network' of all of the grid boxes that meet the data inclusion criteria for 1954–2007, to obtain a best estimate for the most recent period (299 land points, 345 ocean points). The dense network has good spatial coverage between 25° N and 60° N (with some missing values in the interior of Eurasia and at higher latitudes) and more sporadic coverage between 25° S and 40° S. For the comparison network, all of the Southern Ocean, most of the Pacific Ocean and much of Asia are excluded.

Trend distribution testing. Tests for the deviation of distribution means from zero are done using the *t*-test (two-tailed) and confidence intervals are *t*-intervals. The standard deviation for the *t*-test is calculated from the observed distribution and the degrees of freedom are estimated as the effective spatial degrees of freedom (ESDOF) of the time-varying field using the moment-matching method of ref. 25, which they describe as appropriate when testing for the difference of a realization from the mean. This method estimates 21 (λ_{land}), 19 (λ_{ocean}), 12 ($\mathcal{G}_{\mathrm{land}}$) and 20 ($\mathcal{G}_{\mathrm{ocean}}$) ESDOF for λ and $\mathcal{G}$ variability. For the late dense network, the estimates are 29 (λ_{land}), 58 (λ_{ocean}), 12 ($\mathcal{G}_{\mathrm{land}}$) and 47 ($\mathcal{G}_{\mathrm{ocean}}$) ESDOF. For summer temperature field variability we use 15 (land) and 9 (ocean) ESDOF for the comparison network and 17 (land) and 30 (ocean) ESDOF for the dense network. For winter temperature field variability we use 8 (land) and 9 (ocean) ESDOF for the comparison network and 6 (land) and 31 (ocean) ESDOF for the dense network. For seasonal-difference (summer temperature minus winter temperature) hypothesis testing we use ESDOF values calculated from fields of annual mean temperature and we use values of 10 (land) and 9 (ocean) ESDOF for the comparison network and 12 (land) and 22 (ocean) ESDOF for the dense network. Recovered ESDOF estimates are comparable to the observation-based estimates of ref. 23 and are notably smaller than the model-based estimate of ref. 24. For testing the average summer and average winter trend distributions, summer is defined as June, July and August in the Northern Hemisphere and as December, January and February in the Southern Hemisphere. Winter is defined as December, January and February in the Northern Hemisphere and June, July and August in the Southern Hemisphere.

Energy balance model. The one-box conceptual model is a one-atmospheric-layer energy balance model, with a black-body surface and a black-body atmosphere, forced with sinusoidally varying short-wave radiation ($S = S_0\cos(2\pi t)$) characteristic of the annual cycle in radiation at 40° N. We add two complications: (1) to consider sensitivity to atmospheric optical properties, we specify atmospheric short-wave absorptivity ($\mathcal{A} = 0.15$), transmissivity ($\mathcal{T} = 0.6$), and reflectivity ($\mathcal{R} = 0.25$), and calculate the effects of multiple reflections following ref. 34; (2) to consider the effects of increasing long-wave optical depth (τ), we calculate a different atmospheric upward-radiating temperature ($T_{\mathrm{up}} = T_{\mathrm{a}} - \Gamma H(\ln(3\tau/2) - 1)$) and downward-radiating temperature ($T_{\mathrm{down}} = T_{\mathrm{a}} + \Gamma H$), which are related to the interior atmospheric temperature (T_{a}) by the atmospheric height (H) and lapse rate (Γ), following ref. 33 (ch. 5). Surface temperature (T_{s}) tendency is a function of the sum of energy fluxes divided by the thermal mass of the surface (c_{s}). On land, the depth of soil that contributes to the thermal inertia is estimated as the square root of the soil diffusivity times the timescale in question, and for annual timescales we use a depth of 4.7 m in calculating c_{s}.

The surface energy budget is then

$$c_{\mathrm{s}}\frac{\partial T_{\mathrm{s}}}{\partial t} = S\left(\frac{\mathcal{T}(1-\alpha_{\mathrm{s}})}{1-\alpha_{\mathrm{s}}\mathcal{R}}\right) + \sigma T_{\mathrm{down}}^4 - \sigma T_{\mathrm{s}}^4$$

and the atmospheric budget is

$$c_{\mathrm{a}}\frac{\partial T_{\mathrm{a}}}{\partial t} = S\left(\mathcal{A} + \frac{\mathcal{A}\mathcal{T}\alpha_{\mathrm{s}}}{1-\alpha_{\mathrm{s}}\mathcal{R}}\right) + \sigma T_{\mathrm{s}}^4 - \sigma T_{\mathrm{up}}^4 - \sigma T_{\mathrm{down}}^4$$

where α_{s} is the surface albedo. The model is run for parameter values typical for land, and we then perturb these values to estimate sensitivity.

Thermal mass changes are equated with soil moisture changes assuming a soil consisting of 10% inorganic matter, 45% organic matter, 5% unfilled airspace and with a soil water content of 40%. A 13% drop in soil moisture then implies that the soil water content drops to 35%.

51. Cuming, M. J&. Hawkins, B. A. *TERDAT: The FNOC System for Terrain Data Extraction and Processing.* Tech. Rep. Mil Project M-254 (second edition); prepared for USN/FNOC (Meteorology International, 1981).

ARTICLES

The nature of the globular- to fibrous-actin transition

Toshiro Oda[1,2], Mitsusada Iwasa[2], Tomoki Aihara[1], Yuichiro Maéda[2,3] & Akihiro Narita[3]

Actin plays crucial parts in cell motility through a dynamic process driven by polymerization and depolymerization, that is, the globular (G) to fibrous (F) actin transition. Although our knowledge about the actin-based cellular functions and the molecules that regulate the G- to F-actin transition is growing, the structural aspects of the transition remain enigmatic. We created a model of F-actin using X-ray fibre diffraction intensities obtained from well oriented sols of rabbit skeletal muscle F-actin to 3.3 Å in the radial direction and 5.6 Å along the equator. Here we show that the G- to F-actin conformational transition is a simple relative rotation of the two major domains by about 20 degrees. As a result of the domain rotation, the actin molecule in the filament is flat. The flat form is essential for the formation of stable, helical F-actin. Our F-actin structure model provides the basis for understanding actin polymerization as well as its molecular interactions with actin-binding proteins.

Actin was discovered in muscle tissue by Straub in 1942 (ref. 1). The fibrous (F) form of actin is a major component of the thin filament in all muscle tissues. Since then, actin has been found in many eukaryotic cells to be the most abundant cytoskeleton protein, and it performs a broad range of important cellular functions in cell motility as well as in locating and transporting protein complexes in the cell. Actin plays these parts through the dynamic assembly and disassembly of structures such as lamellipodia and filopodia, in a process referred to as actin dynamics[2,3]. Actin exists in a dynamic equilibrium between monomeric G-actin and polymerized F-actin[4]. Actin strongly binds one adenosine nucleotide, and the polymerization involved in the G- to F-actin transition activates the ATPase. The ATPase activity drives actin filament treadmilling[5], in which polymerization at one end and depolymerization at the other occur at the same time. Treadmilling is regulated by actin regulatory proteins, which mediate the actin dynamics.

In 1990 the first crystal structure of G-actin was solved[6]; the F-actin atomic model was also proposed in a back-to-back paper[7]. The model has been widely accepted as an approximate low-resolution structure. However, the conformational changes that occur at the G- to F-actin transition and the accompanying activation of the ATPase have remained elusive. A high-resolution structure is required to understand the detailed interactions between the subunits in the filament and the interactions between F-actin and actin-binding proteins. Here we describe a new high-resolution model for the F-actin structure, based on X-ray fibre diffraction data to 3.3 Å in the radial direction and 5.6 Å along the equator, in contrast to the previous atomic models, which were based on fibre diffraction patterns with low resolution to about 8 Å[7–12]. Our model shows that the essential difference between G-actin and F-actin is the relative rotation of the two major domains by about 20°, which gives the F-actin subunit a flat conformation. A second difference is the conformation of the DNase-I binding loop (D-loop) in subdomain 2, which adopts an open loop conformation. There are no other large-scale differences. The flat conformation is also observed in the polymer of the bacterial MreB actin homologue[13], and thus it seems to be a common characteristic of polymers among actin homologues. The G- to F-actin structural transition can account for a number of previous results about the consequences of modifications and mutations of residues[14,15] as well as the binding of small molecules[16,17] on polymerization. This model also provides a basis for understanding the structural stability of F-actin.

Modelling of F-actin

Our model of the F-actin structure is based on X-ray fibre diffraction analyses. An X-ray fibre diffraction pattern is essentially a section through the cylindrically averaged diffraction pattern from a single fibrous molecule. By analysing the diffraction, the structure can be deduced in favourable cases. In general, this is done by constructing a model and then calculating the expected diffraction pattern; by comparing the calculated and observed diffraction patterns, a better model is eventually arrived at. However, cylindrical averaging leads to a significant loss of information. High resolution is essential for an unambiguous answer. To achieve high resolution, we combined several technical advances: we controlled the filament length by adding gelsolin, prepared well-oriented sols of F-actin[18] by the use of an 18 tesla superconducting magnet (actin filaments are diamagnetic), and recorded X-ray fibre diffraction patterns to 3 Å at the SPring-8 beam lines. From the extracted fibre diffraction data to 3.3 Å in the radial direction and to 5.6 Å along the equator, we constructed the model of the F-actin structure by altering the crystal structure of G-actin by use of the normal modes of actin and a molecular dynamics simulation, while monitoring the *R*-factor-non-fit, an equivalent of the free *R*-factor (Supplementary Fig. 1). Figure 1 shows that the diffraction pattern calculated from our F-actin model (left half of the image), especially the peak positions, is consistent with the observed pattern (right).

Subunit conformation

Actin has a nucleotide-binding cleft enclosed by two major domains (Fig. 2c)[6]. In almost all actin crystal structures, the cleft is closed and the two major domains are in a propeller-like twist with each other. This is characteristic of the G-actin conformation (yellow in Fig. 2a, b; Supplementary Fig. 2). In our F-actin model, we found a previously unknown flat conformation, in which the cleft remains closed

[1]X-ray Structural Analysis Research Team, RIKEN SPring-8 Center, RIKEN Harima Institute, 1-1-1, Kouto, Sayo, Hyogo 679-5148, Japan. [2]ERATO project 'Actin-filament dynamics', Japan Science and Technology Agency (JST), 1-1-1, Kouto, Sayo, Hyogo 679-5148, Japan. [3]Structural Biology Research Center and Division of Biological Science, Graduate School of Science, Nagoya University, Furo, Nagoya 464-8601, Japan.

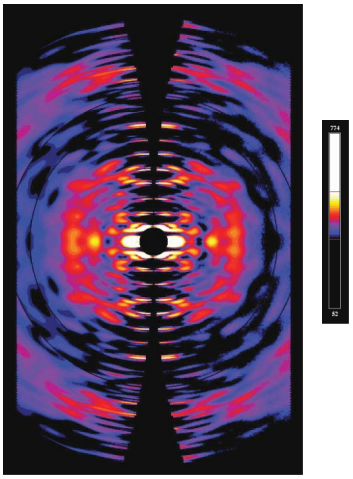


Figure 1 | Comparison between the observed (right half) and calculated (left half) diffraction patterns from our model. The area indicated here, within the outermost boundary of 1/3.3 Å^{-1} in the radial direction and of 1/5.6 Å^{-1} along the equator, was used for the F-actin modelling. The diffraction pattern was calculated by using a disorientation angle of 2.9°, a repeat distance of 9,135 Å and the selection rule $l = -153n + 331m$. The diffraction pattern is divided into three parts (1/60 Å^{-1} and 1/6.5 Å^{-1}, 1/6.5 Å^{-1} and 1/5.5 Å^{-1}, and 1/5.5 Å^{-1} and 1/3.3 Å^{-1}) by the boundaries indicated in black circles, and the scaling was independently performed for each region. The overall similarities between the observed and calculated intensities are satisfactory: *R*-factor-fit = 0.143, *R*-factor-non-fit = 0.207 and *R*-factor-non-fit (overall) = 0.184. The *R*-factor-fit and the *R*-factor-non-fit are defined in Supplementary Fig. 1.

and the two domains are untwisted (cyan in Fig. 2a, b). The two conformations are related by a 20° rotation of the two major domains around the axis passing along the front of subdomain 1 and the side of subdomain 3, which was identified using the program DYNDOM[19] (red line in Fig. 2a, b). The relative rotation of the two major domains is achieved by rotations of the dihedral angles of residues 141–142 and 336–337 (blue in Fig. 2a). Another characteristic of our F-actin subunit is an extended D-loop, to fit the surface of the upper subunit along the strand (Fig. 3b). A 20° relative rotation of the two major domains, together with an extension of the helical D-loop of the tetramethylrhodamine-conjugated actin (TMR–actin) crystal structure[20], would basically account for the peak positions in the low-resolution part of the diffraction pattern, which determine the overall shape of the subunit (*R*-factor fit of ~0.25).

Contacts between subunits

Figure 3 displays the contacts between the subunits within our F-actin model. Each residue with a Cα atom located within 10 Å from any other Cα of a contacting subunit is highlighted in red, blue or yellow. Each subunit is labelled with a number, n, $n+1$ or $n+2$, from the barbed-end side of the filament. The intra-strand interface appears to be extensive (Fig. 3b). The projection 283–294 of subunit $n+2$ (red) is enclosed by residues 61–65, 200–208 and 241–247 of subunit n (blue), resembling a ball (red) in a socket (blue). Additionally, the D-loop 38–49 of subunit n (blue) is extended towards the hydrophobic groove of subunit $n+2$ (red) between subdomains 1 and 3: Val 43 and Met 44 of subunit n contact Leu 346 and Phe 375 of subunit $n+2$.

The inter-strand contacts are formed by two projections from subdomain 4 (Fig. 3c). One is the carboxy terminus of the α-helix 191–199 of subunit n (blue), which contacts the amino terminus of the α-helix 110–115 of subunit $n+1$ (yellow) in the opposite strand. The other is the hydrophobic plug 265–271 in subunit $n+1$ (yellow), which contacts four regions in the opposite strand, including residues 201–203 and 39–42 in subunit n (blue) and residues 170–174 and 285–286 of subunit $n+2$ (red). The contacts mediated by the hydrophobic plug are three-bodied. Residues 39–42, 201–203 and 286 contribute to both the inter- and intra-strand contacts.

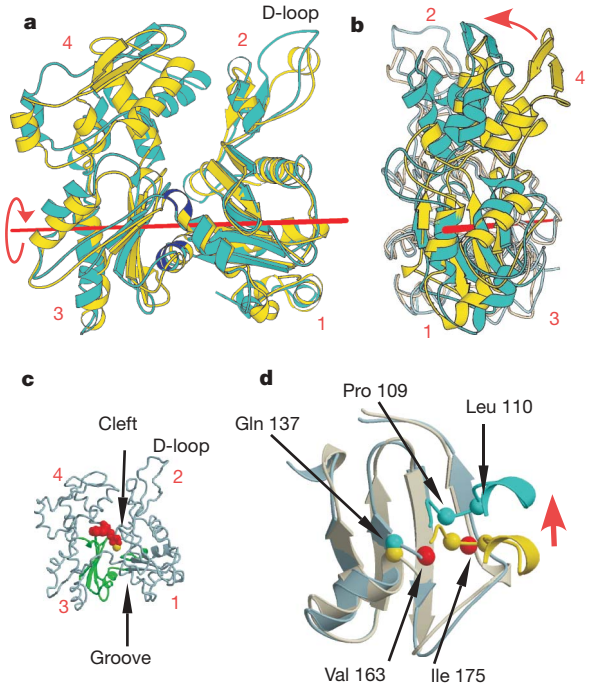


Figure 2 | Transition from the G-actin conformation to the flat conformation in F-actin. **a**, Front view. The subunit in our F-actin model (cyan) and the TMR–actin crystal structure (yellow; PDB code 1J6Z)[20] are superimposed on subdomains 1 and 2. Subdomains 3 and 4 are rotated with respect to subdomains 1 and 2 about the rotational axis (red line) in the direction indicated by the red arrow. The rotation is associated with bends of the polypeptide chain at residues 141–142 and 336–337, as indicated in blue. **b**, Side view, viewed from the left-hand side of subdomains 3 and 4 in **a**. Subdomains 1 and 2 are shown as light cyan and light yellow Cα traces. Stereo views of these panels are shown in Supplementary Fig. 2. **c**, A subunit in our F-actin model, with ADP depicted as red CPK balls and Ca^{2+} as a yellow CPK ball. In **a–c,** the subdomains are labelled with red numbers. **d**, Details of the hinge region indicated in green in **c**. The subunit in our F-actin conformation (cyan) and the TMR–actin crystal structure (yellow) are superimposed on the four-β-strand core of subdomain 3 including Val 163 and Ile 175 (red). The Cα positions are indicated by balls.

Bacterial MreB

Bacterial MreB is a structural and functional homologue of actin[13,21,22]. MreB usually forms thin, non-helical filaments consisting of two protofilaments or bundles of them, and it rarely forms the two-stranded helical polymers seen in F-actin. The MreB crystal contained a straight, single-stranded protofilament, in which the contacts between the subunits were deduced to be biologically relevant[13]. MreB has a closed cleft and the two major domains are untwisted, and thus it resembles the flat conformation of our F-actin subunit (Supplementary Fig. 3a, b). The intra-protofilament contacts are also similar to those of our F-actin model: residues 266–275 are enclosed by three segments, 227–229, 235–238 and 51–59 (Supplementary Fig. 3c). Because the polypeptide chain of actin is longer than that of MreB, there are several insertions within the actin sequence. The actin-specific insertions 197–204 and 264–272 contribute to the inter-strand contacts, which are required to generate the two-stranded helical polymers.

Comparison of our F-actin model with previous models

One important difference between our model and the original Holmes model[7] (and refinements based on this model, the Lorenz model[8]) is the filament diameter. The Holmes model was built to fit the published radius of gyration of 25 Å, which is actually too big. This caused the two strands in the Holmes model to be too far apart,

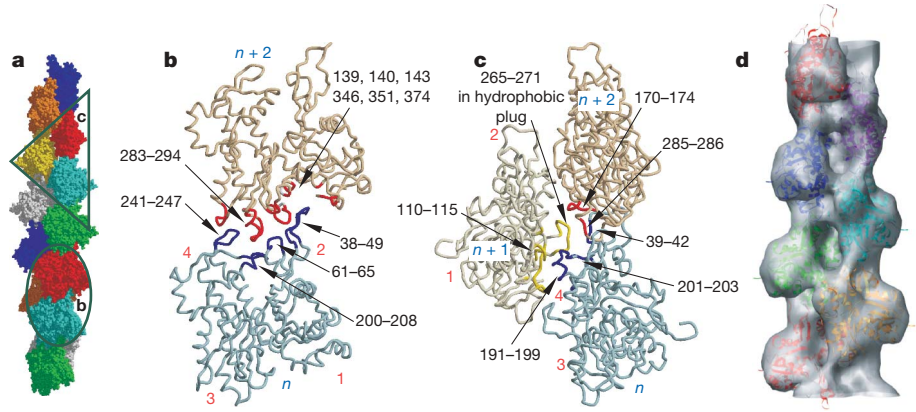


Figure 3 | Intra- and inter-strand contacts within our F-actin model. a, Our model of the F-actin structure including 13 subunits. The two subunits marked by the oval are magnified in **b**, and the three subunits marked by the triangle are magnified in **c**. **b**, **c**, Residues contributing to the intra-strand contacts between subunits (**b**) and those facilitating the inter-strand contacts (**c**) are highlighted. These residues have a Cα within 10 Å from any other Cα of a contacting subunit. Black and red numbers represent the residue and subdomain numbers, respectively, whereas n, $n + 1$ and $n + 2$ are subunit numbers. **d**, The three-dimensional map independently reconstructed from cryoelectron micrographs is superposed with our F-actin model (resolution 13.8 Å). The volume of F-actin is 100%.

which necessitated rebuilding of the hydrophobic plug to achieve contact. Our model is substantially smaller, and has a radius of gyration of 23.7 Å. The smaller filament diameter is clearly substantiated by the three-dimensional map independently reconstructed from cryoelectron micrographs at the resolution of 13.8 Å, which is consistent with our model (Fig. 3d). Holmes *et al.* arrived at a similar conclusion for their actin model, which was built to fit electron micrographs of decorated actin (the Holmes 2003 model)[10]. There seems to be little doubt that the F-actin structure has a smaller diameter than in the original Holmes model. Therefore, it is unlikely that the hydrophobic plug largely alters its conformation upon the G- to F-actin transition because of the narrow inter-strand gap. Nevertheless, in our model the position and the shape of the hydrophobic plug are adjusted within the narrow inter-strand gap, where the plug has major roles as the inter-strand connector. This role is the same as that in the original Holmes model, which is consistent with the results of the actin mutagenesis experiments[23].

Our new model differs from the original Holmes model, and is much closer to the Holmes 2003 model[10]. However, the Holmes 2003 model does not reach the consistency of our model with our independently reconstructed electron micrograph map (Supplementary Figs 4 and 5 and Supplementary Videos 1 and 2). The subunit in the Holmes 2003 model is also flat, but it has a slightly different orientation in the filament and defects in the nucleotide-binding site, probably due to the complex shifts and rotations of each of the four subdomains. This is in contrast to our simple rotation of each of the two major domains. An extra shift of the C terminus applied to the Holmes 2003 model worsens the fitting to our electron micrograph map[24].

Discussion

The previous crystal structures of actin and actin-related proteins are classified into the closed and open conformations, in which the two major domains are twisted relative to each other, regardless of bound nucleotides[20,25–27]. The closed and twisted conformation is the structure of actin that cannot polymerize, and the open and twisted conformation is found with bovine actin-related protein 3 (Arp3) in the inactive Arp2/3 complex[26,28]. The opening of the cleft affects the binding modes of nucleotides[26], and thus the closed-to-open conformational transition is possibly related to nucleotide release. In contrast, the flat conformation of our F-actin subunit constitutes a new class, together with MreB. The flattening seems to be associated with the polymerization and ATPase activity.

The close association between the flattening and the polymerization is supported by the following experimental results. Latrunculin, a toxin isolated from sea sponge, binds to the cleft between the two major domains and inhibits polymerization[17]. The inhibition of polymerization is accounted for by the blocking of the conformational transition due to the toxin binding. In contrast, the nucleotide-free actin polymerizes more readily than the control[29]. The removal of the nucleotide, which connects the two major domains, must facilitate the flattening, thereby promoting the polymerization.

When an actin molecule is incorporated into a filament, the actin ATPase is activated[30]. Gln 137 has a crucial role in the ATPase: the replacement of Gln 137 by alanine markedly reduced the ATPase activity, whereas it promoted the polymerization[31]. The G-actin crystal structures indicated that Gln 137 anchors a water molecule that attacks the γ-phosphate (γ-P) of the bound ATP, and the extremely slow ATPase of G-actin seems to be a consequence of the geometry (the long distance and the shifted orientation) between the water molecule and the γ-P[32]. Gln 137 is located in subdomain 1, whereas ATP tightly binds to subdomains 3 and 4 (Fig. 2c, d). Thus, the domain rotation that is associated with the G- to F-actin transition moves Gln 137 and the γ-P closer to each other, thereby probably allowing the bound ATP to be hydrolysed (Supplementary Fig. 6). In MreB with bound adenylyl imidodiphosphate (AMPPNP), in fact, the γ-P is close (about 4 Å) to the side-chain of the glutamate residue (Glu 131), at the position equivalent to Gln 137 of actin.

However, it should be noted that our F-actin model does not allow us to discuss a more detailed mechanism of the ATPase. This is because the positions of the side chains are not sufficiently accurate (see also Methods Summary) and our model does not include water molecules, which are essential for the ATPase. Therefore, we cannot discuss additional mechanisms that may contribute to the ATPase, such as structural changes in the loop 108–111, as observed in *E. coli* HSP70 homologue Dnak (ref. 33), and a shift of the γ-P through the rearrangement of the two phosphate-binding loops. Moreover, although it seems probable that the flattening of the actin subunit occurs before the ATP hydrolysis, we cannot exclude the possibility that the flattening occurs subsequent to the phosphate release.

The flattening of the actin molecule generates changes in the intramolecular interactions between the two major domains, especially around the two loops that connect them: the 'sensor loop' 71–73 (ref. 25) and residues 108–111 (Supplementary Fig. 7). First, Arg 206, Glu 72, Arg 183, Asp 187, methylated (Me)His 73, Asp 179 and Arg 177 form a concatenation of salt bridges, which probably contribute to stabilizing the F-actin conformation. The highly charged MeHis 73 residue enters the concatenation as a result of the domain rotation, which is probably the cause for MeHis 73 (refs 34 and 35). Second, Pro 109 and Leu 110, in the loop 108–111 of subdomain 1,

detach from the spot around Val 163 and Ile 175 in the hydrophobic core of subdomain 3 (Fig. 2d). This allows Leu 110 to interact with the subunit in the opposite strand at Thr 194, and this subunit–subunit contact also probably stabilizes the flat conformation.

How does the flattening of the actin molecule confer the F-actin formation? A comparison of our F-actin structure with the two-stranded, non-helical straight polymer observed in the crystal of the formin–actin complex[36] provides the answer. In Fig. 4, the intra- and inter-strand contacts of the straight polymer are compared with those of our F-actin model by placing the light-cyan-coloured subunits *n* at the same rotational position about the helix axes. The flattening moves subdomain 4 of the subunit towards the helix axis by about 4 Å, whereas subdomains 1, 2 and 3 are located at about the same radii from the helix axis. As a result, the C terminus of the α-helix (182–196, green) in subdomain 4 of subunit *n* moves to the vicinity of the N terminus of the α-helix (113–126, yellow) in subdomain 1 of subunit $n+1$ (Figs 4a, b), facilitating the inter-strand connections. Furthermore, the flattening generates extensive intra-strand connections. In the straight polymer, the loop 286–289 (red and yellow balls) in the subunit $n+2$ is on residues 200–202 of subunit *n*, and does not contact the loop 242–246 (Fig. 4a). In contrast, in F-actin, the loop 286–289 (red and yellow balls) in subunit $n+2$ is tightly surrounded by three segments, the residues 200–208, the loop 242–246 and the residues 61–65 of subunit *n* (Figs 4b, c). Moreover, the shift of the loop 286–289 enables the residues 202–203 to contact the hydrophobic plug on the opposite strand at Glu 270 of subunit $n+1$. This region is the node between the intra-strand and inter-strand contacts. Thus, the flat conformation allows more extensive contacts between the subunits than in the straight polymer. As expected, filament formation is impaired by substitutions or modifications of the residues involved in these contacts, including the double mutation P243K/A204E (ref. 14), the phosphorylation of Thr 201 to Thr 203 (ref. 37), the mutations I287S and F200S (T.A. and T.O., unpublished data) and the mutations D286R and D288R (Supplementary Fig. 8). The double and triple mutations K61A/R62A, E241A/R244A, D286A/D288A and R290A/K291A/E292A are also lethal in yeast[15].

The previous comparison was made between the non-helical polymer in the formin–actin complex and our F-actin model. Even if we use computer modelling to twist the non-helical polymer into a helical polymer, by azimuthally rotating the G-actin structure, the inter-strand contacts are not strengthened and would remain unstable. Actually, the intra-strand contacts in the computer model are closer to those of the single-stranded, twisted polymer observed in the crystal of actin split by protease from *E. coli* A2 strain (ECP-actin)[38] (Supplementary Fig. 9). The helical disposition alone generates one-half of the intra-strand contacts around residues 286–289 of our F-actin structure, whereas the other half is completed by the additional flattening of the subunits. Thus, stable F-actin is formed by a combination of the flattening of the individual subunits and the helical disposition of the subunits. Nevertheless, the two factors may not necessarily be tightly coupled. According to molecular dynamics simulations, actin favours the closed and twisted conformation regardless of the bound nucleotide[39], and thus the flat conformation in isolation must be less stable. Therefore, the F-actin structure may be influenced by the destabilizing force imposed by the strain of the subunit conformation and the stabilizing effect of the contacts between subunits. We speculate that the balance between the two might induce the multiple local conformations of F-actin, which could account for the dynamics of the F-actin structure, as observed for tubulin and FtsZ[21,40,41].

METHODS SUMMARY

Well-oriented sols of F-actin, which were formed by Ca-actin and length-controlled by gelsolin, were prepared according to the procedures described previously[18]. From the sols thus obtained, we recorded X-ray fibre diffraction patterns at BL41XU, BL40B2 and BL45XU-SAX at SPring-8 (Fig. 1 in Supplementary Methods). From the patterns, the layer-line intensities were extracted and used for modelling of the F-actin structure as 'grouping intensities' re-indexed by the selection rule $l=-6n+13m$. The modelling was started from a rigid body refinement of three crystal structures that were aligned in a straight polymer, as observed in the formin–actin complexes[36]. The resulting models were further refined using 12 elastic normal-mode motions of the actin molecule, the molecular dynamics refinement and the final minimization of the effective energy term with FX-plor[42]. To estimate the restriction of the reflection term imposed on the structure thus obtained, we examined the effects of structural modifications on the *R*-factor. The *R*-factor increased by 5% when a shift of two successive residues by 2 Å was made along the helix axis or a shift of six successive residues by 3 Å was made in the plane perpendicular to the helix axis from the plausible positions. Thus, the reflection term restricts the conformation of the peptide chain at the level of several residues, and the detailed structure is affected by the conformational energy. The structural analysis of F-actin by

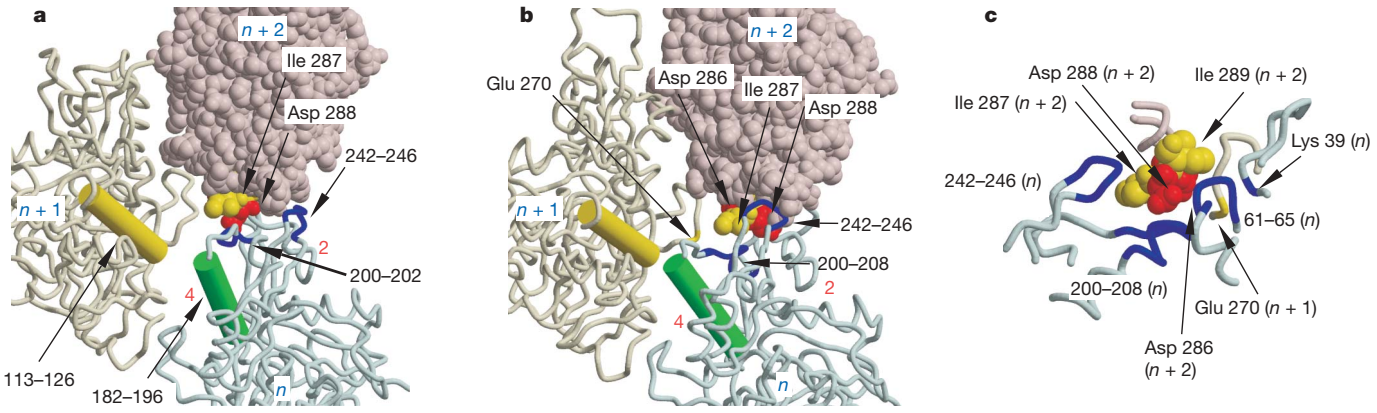


Figure 4 | Comparison between the two-stranded straight polymer in the actin–formin crystal and our helical F-actin polymer. **a**, The two-stranded, non-helical straight polymer observed in the crystal of the actin–formin complex (PDB code 1Y64)[36]. The polymer consists of two protofilaments (the strand including subunit $n+1$ and the strand including subunits *n* and $n+2$) that are straight and parallel to each other. Similar intra-strand contacts are also observed in other crystals[48–50]. **b**, Side view of our F-actin structure. In **a** and **b**, the thick rods are α-helices 113–126 (yellow) of subunit $n+1$ and 182–196 (green) of subunit *n*. **c**, Details of the contacts around Asp 286-Ile 287-Asp 288-Ile 289 of subunit $n+2$ (red and yellow balls), surrounded by three segments of subunit *n* (light cyan). Note that Asp 286 behind Ile 287-Asp 288-Ile 289 joins the network including Arg 39 of subdomain 2, Glu 205 and Thr 203 of subdomain 4, and Glu 270 of the opposite strand. Ile 287 contacts a hydrophobic patch composed of Phe 200, Ala 204, Ile 208 and Pro 243 of subdomain 4. Asp 288 joins the salt bridges including Asp 244, Lys 291 of subdomain 4 and Arg 62 of subdomain 2. Ile 289 contacts Ile 64 of subdomain 2 and Tyr 166 of subunit $n+2$. The residue that interacts with the N terminus of α-helix 202–215 shifts from Asp 288 in the parallel polymer to Asp 286 in our F-actin. In **a**–**c**, the subdomain numbers are indicated in red and the residue numbers are black, and the subunit numbers are shown by *n*, $n+1$ and $n+2$.

cryo-electron microscopy was performed under the same conditions and by the same procedures as those described previously[43,44]. The actin mutagenesis experiment was performed using a baculovirus-based expression system[31]. The protein structure was drawn using Molscript and Raster3D[45–47].

Received 27 March; accepted 28 November 2008.

1. Straub, F. B. in *Studies Int med Chem Univ Szeged* (ed Szent-Györgi), **2,** 3–15 (Karger, 1942).
2. Pollard, T. D. & Borisy, G. G. Cellular motility driven by assembly and disassembly of actin filaments. *Cell* **112,** 453–465 (2003).
3. Carlier, M. F. & Pantaloni, D. Control of actin assembly dynamics in cell motility. *J. Biol. Chem.* **282,** 23005–23009 (2007).
4. Oosawa, F. & Asakura, S. *Thernodynamics of the Polymerization of Protein* (Academic, 1975).
5. Wegner, A. Head to tail polymerization of actin. *J. Mol. Biol.* **108,** 139–150 (1976).
6. Kabsch, W. *et al.* Atomic structure of the actin:DNase I complex. *Nature* **347,** 37–44 (1990).
7. Holmes, K. C., Popp, D., Gebhard, W. & Kabsch, W. Atomic model of the actin filament. *Nature* **347,** 44–49 (1990).
8. Lorenz, M., Popp, D. & Holmes, K. C. Refinement of the F-actin model against X-ray fiber diffraction data by the use of a directed mutation algorithm. *J. Mol. Biol.* **234,** 826–836 (1993).
9. Tirion, M. M., ben-Avraham, D., Lorenz, M. & Holmes, K. C. Normal modes as refinement parameters for the F-actin model. *Biophys. J.* **68,** 5–12 (1995).
10. Holmes, K. C. *et al.* Electron cryo-microscopy shows how strong binding of myosin to actin releases nucleotide. *Nature* **425,** 423–427 (2003).
11. Wu, Y. & Ma, J. Refinement of F-actin model against fiber diffraction data by long-range normal modes. *Biophys. J.* **86,** 116–124 (2004).
12. Oda, T. *et al.* Modeling of the F-actin structure. *Adv. Exp. Med. Biol.* **592,** 385–401 (2007).
13. van den Ent, F., Amos, L. A. & Löwe, J. Prokaryotic origin of the actin cytoskeleton. *Nature* **413,** 39–44 (2001).
14. Joel, P. B., Fagnant, P. M. & Trybus, K. M. Expression of a nonpolymerizable actin mutant in Sf9 cells. *Biochemistry* **43,** 11554–11559 (2004).
15. Wertman, K. F., Drubin, D. G. & Botstein, D. Systematic mutational analysis of the yeast *ACT1* gene. *Genetics* **132,** 337–350 (1992).
16. Allingham, J. S., Klenchin, V. A. & Rayment, I. Actin-targeting natural products: structures, properties and mechanisms of action. *Cell. Mol. Life Sci.* **63,** 2119–2134 (2006).
17. Morton, W. M., Ayscough, K. R. & McLaughlin, P. J. Latrunculin alters the actin–monomer subunit interface to prevent polymerization. *Nature Cell Biol.* **2,** 376–378 (2000).
18. Oda, T. *et al.* Effect of the length and effective diameter of F-actin on the filament orientation in liquid crystalline sols measured by x-ray fiber diffraction. *Biophys. J.* **75,** 2672–2681 (1998).
19. Hayward, S. & Berendsen, H. J. Systematic analysis of domain motions in proteins from conformational change: new results on citrate synthase and T4 lysozyme. *Proteins* **30,** 144–154 (1998).
20. Otterbein, L. R., Graceffa, P. & Dominguez, R. The crystal structure of uncomplexed actin in the ADP state. *Science* **293,** 708–711 (2001).
21. Michie, K. A. & Löwe, J. Dynamic filaments of the bacterial cytoskeleton. *Annu. Rev. Biochem.* **75,** 467–492 (2006).
22. Carballido-Lopez, R. The bacterial actin-like cytoskeleton. *Microbiol. Mol. Biol. Rev.* **70,** 888–909 (2006).
23. Chen, X., Cook, R. K. & Rubenstein, P. A. Yeast actin with a mutation in the "hydrophobic plug" between subdomains 3 and 4 (L266D) displays a cold-sensitive polymerization defect. *J. Cell Biol.* **123,** 1185–1195 (1993).
24. Volkmann, N. *et al.* The structural basis of myosin V processive movement as revealed by electron cryomicroscopy. *Mol. Cell* **19,** 595–605 (2005).
25. Rould, M. A. *et al.* Crystal structures of expressed non-polymerizable monomeric actin in the ADP and ATP states. *J. Biol. Chem.* **281,** 31909–31919 (2006).
26. Nolen, B. J. & Pollard, T. D. Insights into the influence of nucleotides on actin family proteins from seven structures of Arp2/3 complex. *Mol. Cell* **26,** 449–457 (2007).
27. Graceffa, P. & Dominguez, R. Crystal structure of monomeric actin in the ATP state. Structural basis of nucleotide-dependent actin dynamics. *J. Biol. Chem.* **278,** 34172–34180 (2003).
28. Robinson, R. C. *et al.* Crystal structure of Arp2/3 complex. *Science* **294,** 1679–1684 (2001).
29. De La Cruz, E. M. *et al.* Polymerization and structure of nucleotide-free actin filaments. *J. Mol. Biol.* **295,** 517–526 (2000).
30. Pollard, T. D. Regulation of actin filament assembly by Arp2/3 complex and formins. *Annu. Rev. Biophys. Biomol. Struct.* **36,** 451–477 (2007).
31. Iwasa, M. *et al.* Dual roles of Q137 of actin revealed by recombinant human cardiac muscle alpha -actin mutants. *J. Biol. Chem.* **283,** 21045–21053 (2008).
32. Vorobiev, S. *et al.* The structure of nonvertebrate actin: implications for the ATP hydrolytic mechanism. *Proc. Natl Acad. Sci. USA* **100,** 5760–5765 (2003).
33. Vogel, M., Bukau, B. & Mayer, M. P. Allosteric regulation of Hsp70 chaperones by a proline switch. *Mol. Cell* **21,** 359–367 (2006).
34. Yao, X., Nguyen, V., Wriggers, W. & Rubenstein, P. A. Regulation of yeast actin behavior by interaction of charged residues across the interdomain cleft. *J. Biol. Chem.* **277,** 22875–22882 (2002).
35. Nyman, T. *et al.* The role of MeH73 in actin polymerization and ATP hydrolysis. *J. Mol. Biol.* **317,** 577–589 (2002).
36. Otomo, T. *et al.* Structural basis of actin filament nucleation and processive capping by a formin homology 2 domain. *Nature* **433,** 488–494 (2005).
37. Furuhashi, K. *et al.* Phosphorylation by actin kinase of the pointed end domain on the actin molecule. *J. Biol. Chem.* **267,** 9326–9330 (1992).
38. Klenchin, V. A., Khaitlina, S. Y. & Rayment, I. Crystal structure of polymerization-competent actin. *J. Mol. Biol.* **362,** 140–150 (2006).
39. Dalhaimer, P., Pollard, T. D. & Nolen, B. J. Nucleotide-mediated conformational changes of monomeric actin and Arp3 studied by molecular dynamics simulations. *J. Mol. Biol.* **376,** 166–183 (2008).
40. Ravelli, R. B. *et al.* Insight into tubulin regulation from a complex with colchicine and a stathmin-like domain. *Nature* **428,** 198–202 (2004).
41. Wang, H. W. & Nogales, E. Nucleotide-dependent bending flexibility of tubulin regulates microtubule assembly. *Nature* **435,** 911–915 (2005).
42. Wang, H. & Stubbs, G. Molecular dynamics in refinement against fiber diffraction data. *Acta Crystallogr. A* **49,** 504–513 (1993).
43. Narita, A. & Maeda, Y. Molecular determination by electron microscopy of the actin filament end structure. *J. Mol. Biol.* **365,** 480–501 (2007).
44. Narita, A., Takeda, S., Yamashita, A. & Maeda, Y. Structural basis of actin filament capping at the barbed-end: a cryo-electron microscopy study. *EMBO J.* **25,** 5626–5633 (2006).
45. Kraulis, J. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.* **24,** 946–950 (1991).
46. Esnouf, R. M. An extensively modified version of MolScript that includes greatly enhanced coloring capabilities. *J. Mol. Graph. Model.* **15,** 132–134 (1997).
47. Merritt, E. A. & Bacon, D. J. Raster3D: photorealistic molecular graphics. *Methods Enzymol.* **277,** 505–524 (1997).
48. Kudryashov, D. S. *et al.* The crystal structure of a cross-linked actin dimer suggests a detailed molecular interface in F-actin. *Proc. Natl Acad. Sci. USA* **102,** 13105–13110 (2005).
49. Allingham, J. S., Zampella, A., D'Auria, M. V. & Rayment, I. Structures of microfilament destabilizing toxins bound to actin provide insight into toxin design and activity. *Proc. Natl Acad. Sci. USA* **102,** 14527–14532 (2005).
50. Rizvi, S. A., Tereshko, V., Kossiakoff, A. A. & Kozmin, S. A. Structure of bistramide A-actin complex at a 1.35 angstroms resolution. *J. Am. Chem. Soc.* **128,** 3882–3883 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank K. C. Holmes for continuous encouragement and D. Hanein for the gift of the atomic coordinates of the Volkmann *et al.* model (ref. 24). We thank K. Namba, K. Makino and K. Hasegwa for comments on the manuscript, the gift of software package for the fibre analysis and help with recording the diffraction patterns. We also thank S. Fujiwara and K. Mihashi for comments on the manuscript. We finally thank beam-line staffs at SPring-8 BL40B2, BL41XU and BL45XU-SAX, especially M. Kawamoto and K. Ito. The electron microscopy section of this study is partially supported by the Kazato Research Foundation (A.N.).

Author Contributions The X-ray fibre diffraction analysis for F-actin structure was performed by T.O. The mutant analysis of actin was conducted by M.I. and T.A. The structural analysis for the F-actin structure by the use of the electron cryomicroscopy analysis was conducted by A.N. Manuscript preparation was done by T.O. together with Y.M.

Author Information The coordinates for F-actin model have been submitted to PDB under accession number 2ZWH. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to T.O. (toda@spring8.or.jp).

ARTICLES

Intersubunit coordination in a homomeric ring ATPase

Jeffrey R. Moffitt[1]*, Yann R. Chemla[1]*†, K. Aathavan[2], Shelley Grimes[3], Paul J. Jardine[3], Dwight L. Anderson[3,4] & Carlos Bustamante[1,2,5]

Homomeric ring ATPases perform many vital and varied tasks in the cell, ranging from chromosome segregation to protein degradation. Here we report the direct observation of the intersubunit coordination and step size of such a ring ATPase, the double-stranded-DNA packaging motor in the bacteriophage φ29. Using high-resolution optical tweezers, we find that packaging occurs in increments of 10 base pairs (bp). Statistical analysis of the preceding dwell times reveals that multiple ATPs bind during each dwell, and application of high force reveals that these 10-bp increments are composed of four 2.5-bp steps. These results indicate that the hydrolysis cycles of the individual subunits are highly coordinated by means of a mechanism novel for ring ATPases. Furthermore, a step size that is a non-integer number of base pairs demands new models for motor–DNA interactions.

Multimeric ring ATPases of the ASCE (additional strand, conserved E) superfamily represent a structurally homologous yet functionally diverse group of proteins involved in such varied tasks as ATP synthesis, protein unfolding and degradation, and DNA translocation[1–5]. Despite their importance, the coordination mechanism between the hydrolysis cycles of the individual and often identical subunits that compose these ringed proteins is poorly understood. Recent crystallographic and bulk biochemical studies[2,4] suggest various models of coordination in which subunits act sequentially and in order[6–14], simultaneously and in concert[15], or independently and at random[16]. Unfortunately, direct observation of subunit dynamics has only been reported for a heteromeric system, the F1 ring of ATP synthase[8], the heterodimers of which function in a sequential manner.

The DNA packaging motor in the *Bacillus subtilis* bacteriophage φ29 provides a model system to investigate the intersubunit coordination in homomeric ring ATPases because it can be fully reconstituted *in vitro*[17], it has a relatively slow translocation rate[18,19], and it has been extensively characterized by bulk[20] and single-molecule[18,19,21–23] methods. Packaging of the double-stranded (ds)DNA genome of φ29 into its proteinaceous precursor capsid (prohead) is driven by a powerful molecular machine[18] which consists of three multimeric rings organized coaxially around the point of DNA entry[20]: a dodecameric[24] ring of gene product 10 (gp10) known as the head–tail connector; a pentameric[24–26] ring of RNA molecules known as the prohead-RNA (pRNA); and a pentameric[24,26] ring of the ATPase gp16 (see Fig. 1a). Sequence homology[27] places gp16 in the FtsK/HerA family of dsDNA translocases[28]. This family is itself a member of the large ASCE superfamily, thus relating the packaging motor to the ubiquitous AAA+ and RecA-like proteins[3,5].

Recent studies of the packaging motor have suggested a mechanism in which the subunits operate sequentially[19], each binding ATP, hydrolysing it and translocating the DNA by 2 bp[19,29], before the next subunit repeats this cycle. Although this scheme is consistent with the observed data[19,24,30] and with sequential models proposed for other ring ATPases[2,4,6–14], direct observation of the coordination of the mechanochemical cycles of the individual subunits in the packaging motor is still lacking. Here we report the first measurements of the individual packaging steps of the φ29 motor, which reveal both its step size and the novel coordination between its subunits. Because of its relation to the ASCE superfamily[27], the mechanism for the packaging motor we propose here may have implications for the function of a diverse set of ring ATPases.

DNA is packaged in 10-bp increments

To probe the dynamics of the packaging motor of φ29, single prohead–motor–DNA complexes are tethered between two 860-nm-diameter polystyrene beads held in two optical traps as in Fig. 1a. Packaging is initiated *in situ*[22,23] or by restarting stalled complexes[18,19] in an ATP packaging buffer and monitored in a semi-passive mode in which the tension applied to the motor is kept within a narrow range by periodically changing the distance between the two traps (Supplementary Fig. 1). Motor translocation is determined from the decrease in the contour length of the DNA tether and is followed with base-pair-scale resolution[31–33].

In our first experiments we probe packaging at an average low external tension of ~8 pN, and at ATP concentrations ([ATP]) above and below the Michaelis constant (K_m) of the motor, ~30 μM[19]. Figure 1b shows representative packaging traces collected under these conditions. Across the full range of [ATP], packaging of DNA occurs in a stepwise manner consisting of 'dwells', in which the DNA length remains constant, followed by 'bursts', in which DNA is translocated in ~10-bp increments. We determine the average length of DNA encapsidated in these packaging bursts for each [ATP] from the periodicity in the average pairwise distance distribution (PWD) as seen in Fig. 1c. No statistically significant trend is observed in the size of these bursts as a function of [ATP] (see Fig. 1d); thus, the average of these values, 10.0 ± 0.2 bp (s.e.m.), is the best estimate for the burst size.

To elucidate the mechanism by which the motor translocates in 10-bp increments, we analysed the time the motor spends in the dwell

[1]Department of Physics and Jason L. Choy Laboratory of Single Molecule Biophysics, [2]Biophysics Graduate Group, University of California, Berkeley, California 94720, USA. [3]Department of Diagnostic and Biological Sciences, [4]Department of Microbiology, University of Minnesota, Minneapolis, Minnesota 55455, USA. [5]Departments of Molecular and Cell Biology, Chemistry, and Howard Hughes Medical Institute, University of California, Berkeley, California 94720, USA. †Present address: Department of Physics and Center for Biophysics and Computational Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA.
*These authors contributed equally to this work.

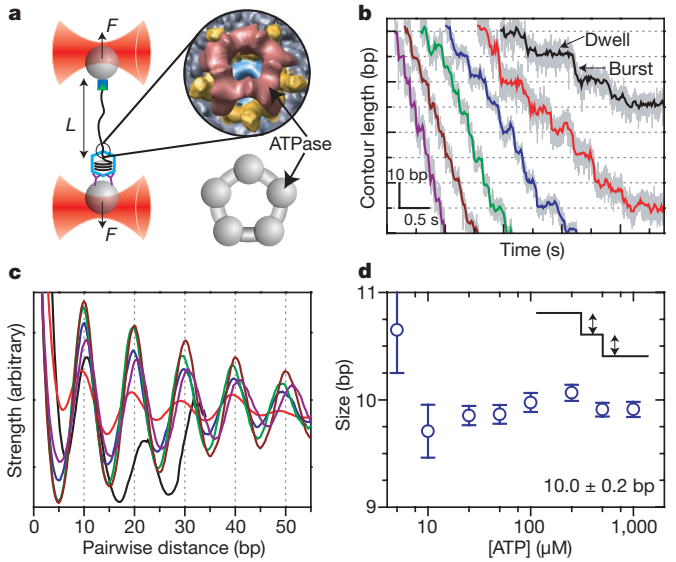


Figure 1 | Bacteriophage φ29 packages DNA in bursts of 10 bp. **a**, A single packaging bacteriophage prohead–motor complex and its dsDNA substrate are tethered between two beads each held in an optical trap and held under tension, *F*. Motor dynamics are inferred from changes in the contour length of the unpackaged DNA, *L*, as a function of time. Inset: cryo-electron microscopy reconstruction of the full motor complex[26] (courtesy of M. Morais), ATPase in red, pRNA in yellow, connector in cyan and capsid in grey with a top view cartoon of the ATPase ring alone (below, grey). **b**, Representative packaging traces collected under low external load, ~8 pN, and different [ATP]: 250 μM, 100 μM, 50 μM, 25 μM, 10 μM and 5 μM in purple, brown, green, blue, red and black, respectively, all boxcar-filtered and decimated to 50 Hz. Data at 1.25 kHz are plotted in light grey. Contour length is plotted in bp of dsDNA. **c**, Average pairwise distributions of packaging traces selected for low noise levels (50% of all packaging data; see Supplementary Figs 2 and 3). Colour scheme as in **b**. **d**, The average size of the packaging burst versus [ATP] determined from the periodicity in **c**. Error bars are the standard deviation in the slope of a linear fit to the peak positions. Data collected at 500 μM and 1 mM [ATP] are not shown in **b** and **c** for clarity.

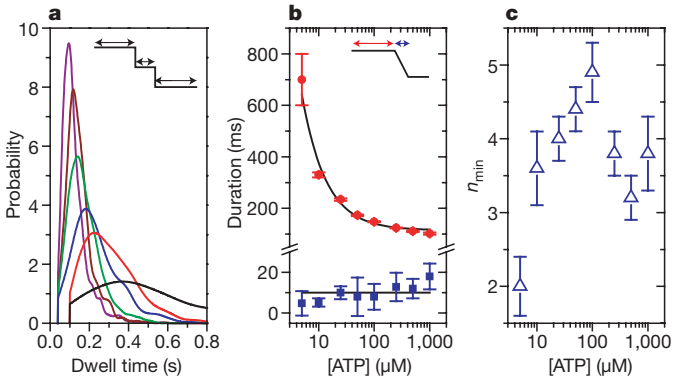


Figure 2 | Dwells before 10-bp bursts contain multiple kinetic events. **a**, Probability distributions for the dwell times preceding a 10-bp burst under low external load, ~8 pN, and different [ATP]: colour scheme as in Fig. 1. Distributions were estimated using kernel density estimation with a Gaussian kernel and the optimum bandwidth[46] and are truncated at the lowest detectable dwell time. Supplementary Fig. 2 contains the number of observed bursts for each [ATP]. Distributions for 500 μM and 1 mM [ATP] are not shown for clarity. **b**, The mean dwell time before the 10-bp bursts (red circles) for all [ATP] with an inverse hyperbolic fit (black line) and the mean duration of all bursts (blue squares, average denoted by black line). **c**, The minimum number of rate-limiting kinetic events during the dwell before the 10-bp bursts, n_{min}, for all [ATP]. Error bars are the standard error.

before each burst and the time it takes to complete each burst as a function of [ATP]. Figure 2a shows the distribution of dwell times before the packaging bursts. The mean dwell time, seen in Fig. 2b, shows a strong dependence on [ATP] that follows an inverse hyperbolic expression, $\langle\tau\rangle = (K_{1/2} + [\mathrm{ATP}])/(k_{max}[\mathrm{ATP}])$, with a $K_{1/2}$ of 23 ± 7 μM (s.d.) and a k_{max} of $8.7 \pm 0.7\ s^{-1}$ (s.d.). In contrast, Fig. 2b shows that the average duration of the packaging burst has little or no dependence on [ATP], suggesting that ATP binding occurs only in the dwells and not in the bursts. Taken together these observations produce a packaging velocity with a Michaelis–Menten [ATP] dependence consistent with previous measurements[19].

The specific shape of the dwell time distributions seen in Fig. 2a provides further information on the kinetic transitions within a single dwell. In particular, the more sharply peaked the distribution, the larger the number of rate-limiting kinetic transitions that compose the dwell[34]. We quantify the degree to which these distributions are peaked with the ratio of the squared mean of the dwell times to their variance, the inverse of the randomness parameter[34] (Fig. 2c). It can be shown that this parameter, n_{min}, provides a strict lower bound[35] on the number of rate-limiting transitions under each [ATP] that occurs during the dwell.

At limiting [ATP], 5 μM $\ll K_m$, we measure an n_{min} of 2.0 ± 0.4 (s.e.m), indicating that there are at least two rate-limiting transitions in each dwell. Because ATP binding must be rate limiting under these conditions, we conclude that no less than two ATP molecules bind to the motor before each 10-bp burst. (In contrast, if a single ATP were to bind during each dwell, one would expect the dwell time distribution to be a single exponential and n_{min} to be 1 (refs 34, 36).) At saturating [ATP], 1 mM $\gg K_m$, we measured an n_{min} of 3.8 ± 0.5 (s.e.m.). Because binding is no longer rate limiting, this indicates that at least four non-binding transitions must also occur in each dwell. For intermediate [ATP], both binding and non-binding transitions can be rate limiting; thus, we expect n_{min} to peak to a value greater than either of the extreme values, exactly as is observed. Thus, Fig. 2c indicates that in total no less than six kinetic transitions must occur in the dwell before each 10-bp burst—at least two ATP binding events and at least four non-binding transitions.

Packaging occurs in four 2.5-bp steps

The findings that packaging occurs in 10-bp increments—five times larger than the 2-bp value proposed from bulk measurements[19,29] —and that the preceding dwells contain multiple ATP binding transitions suggest that the 10-bp bursts may be composed of multiple smaller steps that in general may be too fast to resolve under the above conditions. This inference is supported by the observation that many bursts have durations larger than the measurement bandwidth (Figs 1b and 2b), indicative of intermediate kinetic transitions. Supplementary Fig. 4 shows that occasionally these intermediate transitions can be resolved, appearing as short micro-dwells that split the 10-bp burst into smaller steps. A correlation analysis[36] confirms that these smaller steps occur in groups that sum to 10 bp, ruling out the possibility that these events represent a variable burst size (Supplementary Discussion).

As a direct demonstration of the composition of the 10-bp bursts, we follow packaging against high external loads at near-saturating [ATP] (250 μM). Because translocation steps correspond to force-generating kinetic transitions, we expect that the duration of the micro-dwells preceding these steps will increase with increasing external force[37]. Figure 3a shows that, under 40 pN of average load, smaller steps of ~2.5 bp and integer multiples thereof can be clearly and frequently observed. The PWD for this data, shown in Fig. 3b, reveals a periodicity of 2.4 ± 0.1 bp (s.d.) and the step size distribution, shown in Supplementary Fig. 6, has a peak at 2.48 ± 0.03 bp (s.e.m.). The inset to Fig. 3b shows that the periodicity in the PWD is independent of force, indicating that the 2.5-bp step size is a constant feature of the motor and that the 10-bp bursts observed at low force are composed of four 2.5-bp steps. This conclusion is further supported by the prominent fourth peak observed in the PWD which is consistent with the corresponding 10-bp periodicity observed at low force.

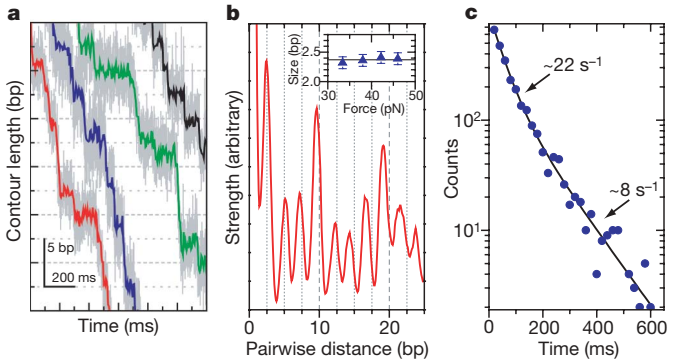


Figure 3 | The 10-bp bursts are composed of four 2.5-bp steps. **a**, Representative packaging traces collected with external loads of ~40 pN and 250 μM [ATP]. Data in light grey are plotted at 1.25 kHz whereas data in colour are boxcar-filtered and decimated to 100 Hz. **b**, Average pairwise distribution of packaging traces selected for low noise levels (50% of all packaging data; see Supplementary Figs 2 and 3). Inset: force dependence of the observed spatial periodicity. The solid line is the mean for all forces, 2.4 ± 0.1 bp (s.e.m.). **c**, Dwell time histogram for the 2.5-bp steps observed under the packaging conditions seen in **a** plotted in blue circles with a bi-exponential fit in black (n = 2,662).

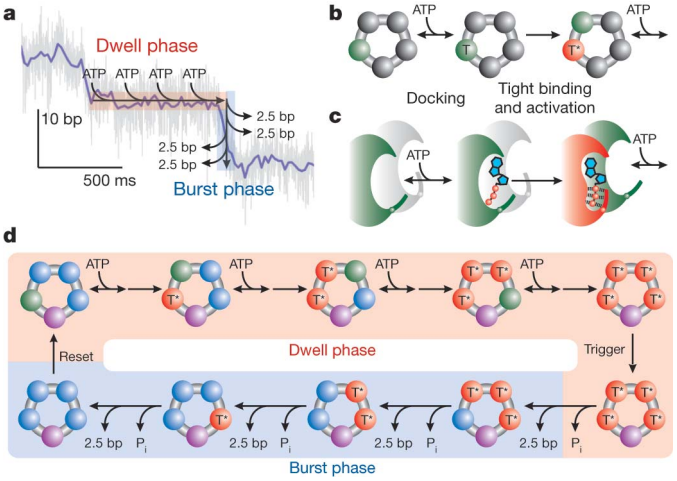


Figure 4 | Intersubunit coordination in the ring ATPase of φ29. **a**, Schematic diagram of the two-phase mechanochemical cycle of φ29 overlaid on a sample packaging trace. **b**, Detailed kinetics of ATP binding. Binding occurs in two steps: ATP docking (green, T) followed by tight binding (red, T*). **c**, Schematic diagram of the communication between subunits during ATP binding. Upon tight binding of an ATP, the binding pocket of the next subunit, formerly inactive (grey), is activated for docking (green). **d**, Schematic depiction of the full mechanochemical cycle of φ29. During the burst phase, ADP may remain on the ring (blue) to be released in the dwell phase. One subunit must be distinct from the others (purple) to break the symmetry of the motor and generate only four steps per cycle. The identity of this subunit may change each cycle. P_i, inorganic phosphate.

The dwell time distribution associated with the 2.5-bp steps (Fig. 3c) is well described by a weighted sum of two exponential decays, with a fast rate of 22 ± 2 s^{-1} (s.d.) and a slow rate of 8 ± 1 s^{-1} (s.d.). The fast rate rationalizes the fraction of 2.5-bp steps that are missed in our analysis and the slow rate is consistent with one out of every four dwells coming from the same peaked dwell time distribution observed at low force (Supplementary Figs 5 and 6). Finally, our data do not support alternative interpretations of the 2.5-bp periodicity, such as distortions from B-form or alternating integer steps, as discussed in Supplementary Figs 7 and 8 and in the Supplementary Information, although some variability in the step size on the ~0.1-bp scale cannot be ruled out.

Intersubunit coordination

Taken together these results indicate that the mechanochemical cycles of the identical subunits of the packaging motor of φ29 are highly coordinated, with the loading of ATP and the translocation of DNA segregated into two distinct phases that comprise the mechanochemical cycle of the entire ring (Fig. 4a). During the dwell phase the DNA is held at constant length while multiple ATPs are loaded, giving this dwell its observed [ATP] dependence (Fig. 2). This process is followed by the burst phase in which DNA is packaged in four increments of 2.5 bp, totalling 10 bp of DNA translocated per cycle (Figs 1d and 3b). Thus, this phase has an average duration that is independent of [ATP] but dependent on force (Fig. 2b).

The observation of four translocation steps per burst strongly suggests that four ATPs bind to the ring during each dwell, one for each of the subsequent steps in the burst phase. This inference is consistent with the measured value of n_{min} at limiting [ATP] as reversibility in binding or differences in binding rates will decrease the observed value of n_{min} from the actual number of binding events[34]. It is also consistent with the 10-bp burst size, as a single ATP provides insufficient free energy to package 10 bp against the high forces tested previously[18,19,37]. Moreover, the binding of four ATPs predicts a coupling constant between ATP consumption and packaging of 2.5 bp per ATP, in reasonable agreement with the ~2 bp per ATP value estimated from bulk measurements[19,29]. The ~25% discrepancy may be explained by additional processes that consume ATP in bulk measurements, such as the repackaging of DNA that slips from the capsid[18,19]. However, it is also possible that a regulatory fifth ATP is bound each cycle and hydrolysed futilely.

Our data also restrict the possible mechanisms by which these ATPs bind to the ring. The requirement that multiple substrate molecules bind per cycle typically results in a sigmoidal dependence on [ATP][38]—the hallmark of binding cooperativity. Yet previous velocity measurements[19] and the mean dwell times measured here (Fig. 2b) are well described by a simple, non-sigmoidal ATP dependence. Because a sigmoidal [ATP] dependence arises whenever two or more binding events are connected reversibly[38], these two observations can be reconciled if and only if the binding of each ATP is separated from the other binding events by a largely irreversible transition. In this case the mean dwell time will display a non-sigmoidal [ATP] dependence despite the cooperative binding of ATP (Supplementary Discussion). This requirement is consistent with previous observations for φ29 (ref. 19) and related ring ATPases[39] which indicate that binding occurs in at least two kinetic steps: (1) a reversible 'docking' transition in which the molecule comes in weak contact with the catalytic pocket; followed by (2) a largely irreversible 'tight-binding' transition[39] in which ATP makes a stronger contact to the binding site and is committed to the hydrolysis cycle (Fig. 4b). More intermediate kinetic states in ATP binding are also possible, but are not required to explain our observations.

In addition, the non-sigmoidal [ATP] dependence of the mean dwell times also restricts the temporal order in which the subunits can dock ATP. Kinetic schemes in which multiple subunits are capable of reversibly docking ATP simultaneously will necessarily have a sigmoidal [ATP] dependence because such schemes have binding events that are reversibly connected[38] (Supplementary Discussion). Thus, it is not sufficient to require that each loose docking of ATP be followed by a tight-binding transition; it is also required that only one subunit at a time can be involved in ATP docking. The simplest model that produces this time-ordered docking is one in which the tight-binding transition of one subunit allosterically activates the binding pocket of another subunit, making it competent to dock ATP, a process depicted in Fig. 4c (Supplementary Discussion). Although our data cannot uniquely determine the actual sequence in which the subunits bind ATP, this required allosteric activation in combination with the known interfacial interactions of adjacent subunits in related ring ATPases[3,5] strongly favours a mechanism in which successive ATP binding occurs in a sequential and ordinal fashion around the ring as depicted in Fig. 4b–d.

Figure 4d summarizes the kinetic transitions that occur during a complete mechanochemical cycle of the packaging motor. During the binding phase, four ATPs bind to the ring in the two-step process depicted in Fig. 4b, c. Previous work has shown that the release of phosphate precedes or coincides with translocation[19]. Thus, after the ring has bound four ATPs, the burst phase is triggered, the first phosphate is released, and the first 2.5-bp step is taken. The burst phase then proceeds with three additional 2.5-bp steps preceded by three force-dependent micro-dwells. The number of rate-limiting steps, n_{min}, at saturating ATP (Fig. 2c) indicates that multiple kinetic transitions in addition to ATP binding must occur during the dwell phase. These transitions may correspond to the hydrolysis of the bound ATPs or the release of multiple ADPs from the previous cycle or, perhaps, both. Moreover, these transitions may occur together either as trigger or reset processes (Fig. 4d) or interspersed between ATP binding events. It is also possible that these additional events correspond to the tight-binding transitions, although this is unlikely given that tight binding is believed to occur quite rapidly[19].

The two-phase model we propose here is also consistent with previous measurements of the packaging motor[18,19]. For example, it has been shown that the binding of a single non-hydrolysable ATP analogue is sufficient to pause the entire motor[19]—a result consistent with the high degree of intersubunit coordination observed here. Furthermore, a biphasic sedimentation profile observed in sucrose gradient experiments suggests the ability of the ring to load multiple nucleotides[19], consistent with our model. Finally, the two-phase model predicts the same dependence of the packaging velocity with force and [ATP] as observed previously[18,19].

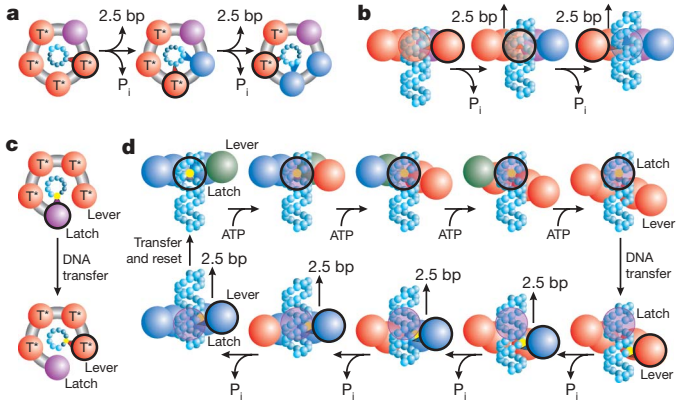


Figure 5 | Packaging models that produce a non-integer step size. **a**, Depiction of a translocation model in which all subunits eventually contact the DNA (cyan spheres). The contacting subunit is outlined in black (top view). **b**, In such a model the size of internal conformational changes set the step size (side view). **c**, Depiction of a translocation model in which only two subunits contact the DNA (black outline). **d**, In such a model, one subunit maintains contact with the DNA (the latch) while the loading of each ATP introduces relative subunit–subunit rotations which distort the ring. This distortion extends one subunit (the lever) along the DNA by ~10 bp. The DNA contact point is then transferred from the latch to the lever, and the release of hydrolysis products relaxes the ring, retracting the lever and the DNA. The DNA contact is then transferred back to the latch, the ring resets and the cycle begins again. Because there are four subunits, the ring is retracted in four steps, dividing a 10-bp step into four ~2.5-bp substeps. The subunit colour scheme is the same as in Fig. 4.

Non-integer step-size models

Our finding that packaging occurs in four 2.5-bp translocation steps raises two notable questions on the motor mechanism. First, how does a dsDNA translocase move in a non-integer number of base pairs? And, second, how is the pentameric symmetry[24–26] of the motor broken, generating only four steps per cycle? A step size that is a non-integer number of base pairs prohibits any mechanism in which every motor subunit within a closed ring makes specific and identical chemical contacts with one strand of the DNA. Under this constraint, we can speculate on several alternative mechanisms that would produce a 2.5-bp step size and the implications these models have for a pentameric motor.

A non-integer step size could be generated if each subunit is capable of binding two or more alternating chemical moieties, which may or may not be on the same strand. Alternatively, the motor may make no specific contacts, but rather drives translocation by means of steric interactions, in which case the step size would be set not by the chemical periodicity of the DNA but by the size of the internal conformational changes that generate the power stroke. One example of such a mechanism is depicted in Fig. 5a, b where each subunit makes non-specific contacts with the major groove of the DNA. In such a model, generation of four translocation steps requires that one of the five subunits is not equivalent to the other four, breaking the symmetry of the pentameric ring. Because the nucleotide-free state is disengaged from the DNA[19], one subunit may be required to retain nucleotide at the end of the previous cycle, ensuring that a strong contact with the DNA is maintained while the remaining subunits load ATP during the subsequent dwell phase.

Alternatively, a single specific chemical contact may be made with the DNA but not with every subunit. In this class of models, only a subset of the subunits interacts with the DNA and relative motion between these subunits is what drives translocation. Figure 5c, d depicts an example of such a mechanism in which only two subunits make specific contact with the DNA. Translocation is achieved via an 'inchworm-like' movement of these two subunits driven by distortions in the ring. One appeal of this mechanism is that because a single specific contact is made with the DNA, it produces an integer burst size, yet because the DNA-binding subunits are retracted by conformational changes induced into the ring by the other subunits, this burst can be divided into non-integer steps. Moreover, this model also explains naturally the observation of four steps by a pentameric motor, as one subunit interface must bear the accumulated distortion of the other four subunits, perhaps inactivating one of the five binding pockets. The relative motion between adjacent subunits needed to accommodate such a mechanism has been observed in the crystal structures of other ring ATPases[6,40] but has not been implicated as part of the translocation mechanism[41]. Future measurements will be aimed at testing the spectrum of models discussed here.

Conclusions

We have presented here the first high-resolution measurements of the stepping dynamics of the ring ATPase of the packaging motor of bacteriophage φ29. Our results indicate a highly coordinated two-phase mechanism in which the binding of ATP and the translocation of DNA by multiple subunits are organized into two distinct and temporally segregated portions of the mechanochemical cycle of the ring. Our observation of a 2.5-bp step size challenges the long-held view that DNA translocation must occur in integer base-pair increments, making it necessary to devise new and more complex models for motor–DNA interactions. In addition, although the inter-subunit coordination we observe is reminiscent of aspects of both the concerted-action model of the large tumour antigen of SV40 (ref. 15) and the sequential models proposed for the translocases BPV E1, T7 gp4, ϕ12 P4, *Escherichia coli* Rho and FtsK[6,7,9–14], our mechanism represents a novel type of coordination not previously proposed for ring ATPases. Provocatively, although a two-phase mechanism contrasts with these other models, it seems to be consistent with many of the biochemical[11–14] and structural[6,7,9,10,13] observations made on these related systems. One notable exception is the ClpX protease for which biochemical data clearly suggest a limited degree of subunit coordination[16]. However, recent work on a related system, the archaeal MCM, suggests that coordinated systems can take alternative pathways when overcoming functional barriers such as catalytically inactive subunits[42]. Ring ATPases of the ASCE superfamily support a large and remarkably diverse set of cellular functions by

drawing on a comparatively small set of common structural features. Direct measurements of the intersubunit dynamics in these systems, such as those presented here, promise to reveal if these diverse cellular functions arise from a similarly small set of common structural dynamics.

METHODS SUMMARY

Complexes of prohead, gp16 and biotinylated DNA were prepared and attached to 860-nm-diameter polystyrene beads (Spherotech) coated with antibodies to φ29 or streptavidin using methods that have been described previously[18,19]. Tethers were assembled and packaging was restarted in a packaging buffer (50 mM Tris-HCl, 50 mM NaCl, 5 mM $MgCl_2$, 10 μg ml^{-1} BSA, 0.1% NaN_3, pH 7.8) supplemented with various amounts of ATP (Sigma-Aldrich)[18,19]. Experiments were conducted in two separate dual-trap instruments, built around two different trapping lasers[32,33], and calibrated using standard techniques[32,33,43]. The contour length of the DNA tether was calculated from the measured force and extension using the extensible worm-like-chain model as described previously[18,19]. Pairwise distributions were calculated from data filtered with a sliding 20-ms boxcar window as described previously[44]. The location and duration of stepping transitions were found with a *t*-test analysis similar to previous methods[45]. Dwell times were calculated directly from the time between transitions, and burst durations were calculated from the number of points within transitions. The mean and variance were calculated directly from these dwell times, and the errors were estimated using a bootstrap method. n_{min} was calculated directly from these moments[34,36].

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 14 May; accepted 11 November 2008.
Published online 7 January 2009.

1. Latterich, M. & Patel, S. The AAA team: related ATPases with diverse functions. *Trends Cell Biol.* **8,** 65–71 (1998).
2. Ogura, T. & Wilkinson, A. J. AAA+ superfamily ATPases: common structure–diverse function. *Genes Cells* **6,** 575–597 (2001).
3. Iyer, L. M., Leipe, D. D., Koonin, E. V. & Aravind, L. Evolutionary history and higher order classification of AAA+ ATPases. *J. Struct. Biol.* **146,** 11–31 (2004).
4. Kainov, D. E., Tuma, R. & Mancini, E. J. Hexameric molecular motors: P4 packaging ATPase unravels the mechanism. *Cell. Mol. Life Sci.* **63,** 1095–1105 (2006).
5. Erzberger, J. P. & Berger, J. M. Evolutionary relationships and structural mechanisms of AAA+ proteins. *Annu. Rev. Biophys. Biomol. Struct.* **35,** 93–114 (2006).
6. Singleton, M. R., Sawaya, M. R., Ellenberger, T. & Wigley, D. B. Crystal structure of T7 gene 4 ring helicase indicates a mechanism for sequential hydrolysis of nucleotides. *Cell* **101,** 589–600 (2000).
7. Mancini, E. J. *et al.* Atomic snapshots of an RNA packaging motor reveal conformational changes linking ATP hydrolysis to RNA translocation. *Cell* **118,** 743–755 (2004).
8. Kinosita, K., Adachi, K. & Itoh, H. Rotation of F1-ATPase: How an ATP-driven molecular machine may work. *Annu. Rev. Biophys. Biomol. Struct.* **33,** 245–268 (2004).
9. Enemark, E. J. & Joshua-Tor, L. Mechanism of DNA translocation in a replicative hexameric helicase. *Nature* **442,** 270–275 (2006).
10. Skordalakes, E. & Berger, J. M. Structural insights into RNA-dependent ring closure and ATPase activation by the Rho termination factor. *Cell* **127,** 553–564 (2006).
11. Adelman, J. L. *et al.* Mechanochemistry of transcription termination factor Rho. *Mol. Cell* **22,** 611–621 (2006).
12. Liao, J.-C., Jeong, Y.-J., Kim, D.-E., Patel, S. S. & Oster, G. Mechanochemistry of T7 DNA helicase. *J. Mol. Biol.* **350,** 452–475 (2005).
13. Massey, T. H., Mercogliano, C. P., Yates, J., Sherratt, D. J. & Löwe, J. Double-stranded DNA translocation: structure and mechanism of hexameric FtsK. *Mol. Cell* **23,** 457–469 (2006).
14. Crampton, D. J., Mukherjee, S. & Richardson, C. C. DNA-induced switch from independent to sequential dTTP hydrolysis in the bacteriophage T7 DNA helicase. *Mol. Cell* **21,** 165–174 (2006).
15. Gai, D., Zhao, R., Li, D., Finkielstein, C. V. & Chen, X. S. Mechanisms of conformational change for a replicative hexameric helicase of SV40 large tumor antigen. *Cell* **119,** 47–60 (2004).
16. Martin, A., Baker, T. A. & Sauer, R. T. Rebuilt AAA+ motors reveal operating principles for ATP-fuelled machines. *Nature* **437,** 1115–1120 (2005).
17. Guo, P., Grimes, S. & Anderson, D. A defined system for *in vitro* packaging of DNA-gp3 of the *Bacillus subtilis* bacteriophage φ29. *Proc. Natl Acad. Sci. USA* **83,** 3505–3509 (1986).
18. Smith, D. E. *et al.* The bacteriophage straight φ29 portal motor can package DNA against a large internal force. *Nature* **413,** 748–752 (2001).
19. Chemla, Y. R. *et al.* Mechanism of force generation of a viral DNA packaging motor. *Cell* **122,** 683–692 (2005).
20. Grimes, S., Jardine, P. J. & Anderson, D. Bacteriophage φ29 DNA packaging. *Adv. Virus Res.* **58,** 255–294 (2002).
21. Hugel, T. *et al.* Experimental test of connector rotation during DNA packaging into bacteriophage φ29 capsids. *PLoS Biol.* **5,** e59 (2007).
22. Fuller, D. N. *et al.* Ionic effects on viral DNA packaging and portal motor function in bacteriophage φ29. *Proc. Natl Acad. Sci. USA* **104,** 11245–11250 (2007).
23. Rickgauer, J. P. *et al.* Portal motor velocity and internal force resisting viral DNA packaging in bacteriophage φ29. *Biophys. J.* **94,** 159–167 (2008).
24. Simpson, A. A. *et al.* Structure of the bacteriophage φ29 DNA packaging motor. *Nature* **408,** 745–750 (2000).
25. Morais, M. C. *et al.* Cryoelectron-microscopy image reconstruction of symmetry mismatches in bacteriophage φ29. *J. Struct. Biol.* **135,** 38–46 (2001).
26. Morais, M. C. *et al.* Defining molecular and domain boundaries in the bacteriophage φ29 DNA packaging motor. *Structure* **16,** 1267–1274 (2008).
27. Burroughs, A. M., Iyer, L. M. & Aravind, L. in *Gene and Protein Evolution* (ed. Volff, J.-N.) 48–65 (Karger, 2007).
28. Iyer, L. M., Makarova, K. S., Koonin, E. V. & Aravind, L. Comparative genomics of the FtsK-HerA superfamily of pumping ATPases: implications for the origins of chromosome segregation, cell division and viral capsid packaging. *Nucleic Acids Res.* **32,** 5260–5279 (2004).
29. Guo, P., Peterson, C. & Anderson, D. Prohead and DNA-gp3-dependent ATPase activity of the DNA packaging protein gp16 of bacteriophage φ29. *J. Mol. Biol.* **197,** 229–236 (1987).
30. Chen, C. & Guo, P. Sequential action of six virus-encoded DNA-packaging RNAs during phage φ29 genomic DNA translocation. *J. Virol.* **71,** 3864–3871 (1997).
31. Moffitt, J. R., Chemla, Y. R., Smith, S. B. & Bustamante, C. Recent advances in optical tweezers. *Annu. Rev. Biochem.* **77,** 205–228 (2008).
32. Moffitt, J. R., Chemla, Y. R., Izhaky, D. & Bustamante, C. Differential detection of dual traps improves the spatial resolution of optical tweezers. *Proc. Natl Acad. Sci. USA* **103,** 9006–9011 (2006).
33. Bustamante, C., Chemla, Y. R. & Moffitt, J. R. in *Single-Molecule Techniques: A Laboratory Manual* (eds Selvin, P. R. & Ha, T.) 297–324 (Cold Spring Harbor Laboratories, 2008).
34. Schnitzer, M. J. & Block, S. M. Statistical kinetics of processive enzymes. *Cold Spring Harb. Symp. Quant. Biol.* **60,** 793–802 (1995).
35. Koza, Z. Maximal force exerted by a molecular motor. *Phys. Rev. E* **65,** 031905 (2002).
36. Chemla, Y. R., Moffitt, J. R. & Bustamante, C. Exact solutions for kinetic models of macromolecular dynamics. *J. Phys. Chem. B* **112,** 6025–6044 (2008).
37. Bustamante, C., Chemla, Y. R., Forde, N. R. & Izhaky, D. Mechanical processes in biochemistry. *Annu. Rev. Biochem.* **73,** 705–748 (2004).
38. Segel, I. H. *Enzyme Kinetics* (John Wiley & Sons, 1975).
39. Oster, G. & Wang, H. Reverse engineering a protein: the mechanochemistry of ATP synthase. *Biochim. Biophys. Acta* **1458,** 482–510 (2000).
40. Skordalakes, E. & Berger, J. M. Structure of the Rho transcription terminator: mechanism of mRNA recognition and helicase loading. *Cell* **114,** 135–146 (2003).
41. Lisal, J. *et al.* Functional visualization of viral molecular motor by hydrogen-deuterium exchange reveals transient states. *Nature Struct. Mol. Biol.* **12,** 460–466 (2005).
42. Moreau, M. J., McGeoch, A. T., Lowe, A. R., Itzhaki, L. S. & Bell, S. D. ATPase site architecture and helicase mechanism of an archaeal MCM. *Mol. Cell* **28,** 304–314 (2007).
43. Berg-Sorensen, K. & Flyvbjerg, H. Power spectrum analysis for optical tweezers. *Rev. Sci. Instrum.* **75,** 594–612 (2004).
44. Block, S. M. & Svoboda, K. Analysis of high resolution recordings of motor movement. *Biophys. J.* **68,** 230–241 (1995).
45. Carter, N. J. & Cross, R. A. Mechanics of the kinesin step. *Nature* **435,** 308–312 (2005).
46. Parzen, E. On estimation of a probability density function and mode. *Ann. Math. Stat.* **33,** 1065–1076 (1962).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank C. L. Hetherington, M. Nollmann and G. Chistol for a critical reading of the manuscript; C. L. Hetherington, A. Politzer, M. Strycharska, M. Kopaczynska and J. Yu for critical discussions; and J. Choy, S. Grill and S. Smith for advice regarding instrumentation. J.R.M. acknowledges the National Science Foundation's Graduate Research Fellowship and Y.R.C. the Burroughs Welcome Fund's Career Awards at the Scientific Interface for funding. This research was supported in part by NIH grants GM-071552, DE-003606 and GM-059604. The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author Contributions J.R.M., Y.R.C. and K.A. conducted the experiments and performed the analysis; S.G., P.J.J. and D.L.A. prepared and provided experimental materials; and J.R.M., Y.R.C., K.A., S.G., P.J.J., D.L.A. and C.B. wrote the paper. J.R.M. and Y.R.C. contributed equally to this work.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to C.B. (carlos@alice.berkeley.edu).

METHODS

Sample preparation. Proheads, gp16 and genomic DNA were isolated as described previously[47]. For a stalled-complex method of initiation[18,19], a ClaI (New England Biolabs) digest of genomic DNA was biotinylated using Klenow exo$^-$ (New England Biolabs) to fill in the overhang with biotinylated nucleotides (Sigma-Aldrich)[18,19]. Preferential packaging of the left end of the genome[20] favours the formation of prohead–motor–DNA complexes with the 6,149-bp fragment of the ClaI digest. Stalled complexes were then bound to antibody beads, made as described previously[18], and introduced into the tweezers with streptavidin-coated beads. For the *in situ* method of initiation[22,23], a 4,277-bp tether PCR amplified from lambda DNA with a biotinylated primer was bound to streptavidin-coated beads, and stalled prohead–motor complexes[22,23] were bound to antibody beads. In both initiation methods, tethers were formed in the tweezers by physically bumping the two beads. The *in situ* initiation method was used for all [ATP] $\geq$ 25 μM as data were in general less noisy and easier to collect; however, a severe drop in tether formation efficiency below 25 μM required the use of the stalled complex method for low [ATP]. All tether lengths were selected to reduce the effect of packaged DNA on motor dynamics[18,23].

Optical trapping instruments. Two different optical trapping instruments were used in these studies[32,33]. All low-force data for 25 μM $\leq$ [ATP] $\leq$ 250 μM were collected using an instrument built around a 845-nm, 200-mW diode laser[32]. All other data were collected using an instrument built around a high-power, diode-pumped, solid-state 1,064-nm laser[33]. Both instruments exploit the correlations in the motion of the two trapped beads with a differential detection technique[32] to achieve base-pair resolution on the second timescale[31–33]. Owing to increased laser absorption at 1,064 nm[48], the high-force data were collected in an 80% deuterium-oxide (D_2O) buffer to avoid heating effects caused by the high laser power needed to provide the large opposing forces. Supplementary Fig. 5 shows that although D_2O alters the kinetics of packaging, it does not change the size of the packaging bursts. In addition, when working with the 1,064-nm trapping laser, an oxygen scavenging system was added (100 μg ml^{-1} glucose oxidase, 20 μg ml^{-1} catalase, 5 mg ml^{-1} dextrose; Sigma-Aldrich) to prevent the formation of the reactive species singlet oxygen.

Calibration. Traps were calibrated using the thermal fluctuations of the trapped beads[43]. The contour length was calculated from the measured extension and force with the extensible worm-like-chain model using a persistence length of 53 nm, a stretch modulus of 1,200 pN[19,49], and an average B-form DNA rise of 3.4 Å bp^{-1} (ref. 50). Distance calibrations were corroborated with video microscopy[33], which was calibrated to 0.3% with two different distance standards (Nikon; Graticules) and confirmed to 1% by measuring the extension of DNA of different lengths, ~1, 2, 3 and 5.6 kb. All packaging experiments were conducted in a semi-passive mode (Supplementary Fig. 1), in which the trap separation was kept constant as packaging proceeded and was changed discretely to keep the tension within a set range: ~6–10 pN for low-force experiments and ~33–46 pN for high-force experiments. All reported data have been corrected for small systematic errors, ~4% (Supplementary Figs 1 and 2), determined from the discrete changes in the trap separation as described in the Supplementary Discussion.

Analysis. The one-sided autocorrelation of a positional histogram of each semi-passive mode segment was used to calculate the pairwise distributions[44]. 0.25-bp bins and 0.1-bp bins were used for the histograms for the low-force and high-force distributions, respectively. Data were selected for low noise and clarity of steps as described in the Supplementary Discussion. The pairwise distributions of the selected data were averaged together to produce the reported distributions. The average spatial periodicity was quantified from the position of the peaks. Subsets of the high force data were analysed to produce the step-size measurement as a function of force.

Stepping transitions were identified using a *t*-test analysis[45] and a probability threshold of observing a given *t*-value of 10^{-4}. Dwell times were calculated from the time between transitions and step sizes were calculated from the difference in mean position between transitions. The reported dwell time distributions in Figs 2a and 3c were selected based on the size of the subsequent step: 8–12 bp and 1.5–4 bp, respectively. Burst durations were estimated from the exponential decay rate of the distribution of the number of contiguous points for which the *t*-value probability was below the 10^{-4} threshold. Because of our limited time resolution, we could not observe the expected peak in this distribution; thus, we report the observed exponential decay, a value less biased by the time resolution than the mean. The effective bandwidth of the *t*-test algorithm was varied to maximize the number of observed steps while minimizing the systematic errors introduced into the moments of the distribution from a finite dead time. By assuming a Poisson distribution, it was determined that this dead time introduces negligible systematic errors in the calculated moments for all reported distributions.

47. Grimes, S. & Anderson, D. The bacteriophage φ29 packaging proteins supercoil the DNA ends. *J. Mol. Biol.* **266,** 901–914 (1997).
48. Kellner, L. The near infra-red absorption spectrum of heavy water. *Proc. R. Soc. Lond. A* **159,** 0410–0415 (1937).
49. Baumann, C. G., Smith, S. B., Bloomfield, V. A. & Bustamante, C. Ionic effects on the elasticity of single DNA molecules. *Proc. Natl Acad. Sci. USA* **94,** 6185–6190 (1997).
50. Yanagi, K., Prive, G. G. & Dickerson, R. E. Analysis of local helix geometry in three B-DNA decamers and eight dodecamers. *J. Mol. Biol.* **217,** 201–214 (1991).

LETTERS

Cold streams in early massive hot haloes as the main mode of galaxy formation

A. Dekel[1], Y. Birnboim[1,2], G. Engel[1], J. Freundlich[1,3], T. Goerdt[1], M. Mumcuoglu[1], E. Neistein[1,4], C. Pichon[5], R. Teyssier[6,7] & E. Zinger[1]

Massive galaxies in the young Universe, ten billion years ago, formed stars at surprising intensities[1,2]. Although this is commonly attributed to violent mergers, the properties of many of these galaxies are incompatible with such events, showing gas-rich, clumpy, extended rotating disks not dominated by spheroids[1–5]. Cosmological simulations[6] and clustering theory[6,7] are used to explore how these galaxies acquired their gas. Here we report that they are 'stream-fed galaxies', formed from steady, narrow, cold gas streams that penetrate the shock-heated media of massive dark matter haloes[8,9]. A comparison with the observed abundance of star-forming galaxies implies that most of the input gas must rapidly convert to stars. One-third of the stream mass is in gas clumps leading to mergers of mass ratio greater than 1:10, and the rest is in smoother flows. With a merger duty cycle of 0.1, three-quarters of the galaxies forming stars at a given rate are fed by smooth streams. The rarer, submillimetre galaxies that form stars even more intensely[2,12,13] are largely merger-induced starbursts. Unlike destructive mergers, the streams are likely to keep the rotating disk configuration intact, although turbulent and broken into giant star-forming clumps that merge into a central spheroid[4,10,11]. This stream-driven scenario for the formation of discs and spheroids is an alternative to the merger picture.

It appears that the most effective star formers in the Universe were galaxies of stellar and gas masses of $\sim 10^{11} M_{\odot}$ at redshifts $z = 2-3$, when the Universe was ~3 Gyr old. ($M_{\odot}$, solar mass.) The common cases[1,3] show star-formation rates (SFRs) of $100 M_{\odot} - 200 M_{\odot}\,\mathrm{yr}^{-1}$. These include ultraviolet-selected galaxies termed BX/BM galaxies (ref. 14) and rest-frame optically selected galaxies termed sBzK galaxies (ref. 15), to be referred to collectively as 'star-forming galaxies' (SFGs). Their SFRs are much higher than the $4 M_{\odot}\,\mathrm{yr}^{-1}$ in today's Milky Way, although their masses and dynamical times are comparable. The co-moving space density of SFGs is $n \approx 2 \times 10^{-4}\,\mathrm{Mpc}^{-3}$, implying, within the standard cosmology (termed ΛCDM), that they reside in dark matter haloes of mass $\lesssim 3.5 \times 10^{12} M_{\odot}$. The most extreme star formers are dusty submillimetre galaxies (SMG)[12,13], with SFRs of up to $\sim 1{,}000 M_{\odot}\,\mathrm{yr}^{-1}$ and $n \approx 2 \times 10^{-5}\,\mathrm{Mpc}^{-3}$. Whereas most SMGs could be starbursts induced by major mergers, the kinematics of the SFGs indicate extended, clumpy, thick rotating disks that are incompatible with the expected compact or highly perturbed kinematics of ongoing mergers[1,3,4]. The puzzle is how massive galaxies form most of their stars so efficiently at early times through a process other than a major merger. A necessary condition is a steady, rapid gas supply into massive disks.

It is first necessary to verify that the required rate of gas supply is compatible with the cosmological growth rate of dark matter haloes. The average growth rate of halo mass, M_{v}, through mergers and smooth accretion, is derived[6] on the basis of the extended Press–Schechter (EPS) theory of gravitational clustering (Supplementary Information, section 1) or from cosmological simulations[16,17]. For ΛCDM, the corresponding growth rate of the baryonic component is approximately

$$\dot{M} \approx 6.6 M_{12}^{1.15} (1+z)^{2.25} f_{0.165} M_{\odot}\,\mathrm{yr}^{-1} \quad (1)$$

where $M_{12} \equiv M_{\mathrm{v}}/10^{12} M_{\odot}$ and $f_{0.165}$ is the baryonic fraction in the haloes in units of the cosmological value, $f_{\mathrm{b}} = 0.165$. Thus, at $z = 2.2$, the baryonic growth rate of haloes of mass $2 \times 10^{12} M_{\odot}$ is $\dot{M} \approx 200 M_{\odot}\,\mathrm{yr}^{-1}$, sufficient to maintain the SFR in SFGs. However, the margin by which this is sufficient is not large, implying that (1) the incoming material must be mostly gaseous, (2) the cold gas must efficiently penetrate into the inner halo and (3) the SFR must closely follow the gas supply rate.

The deep penetration is not a trivial matter, given that halo masses of $M_{\mathrm{v}} > 10^{12} M_{\odot}$ are above the threshold for virial shock heating[8,9,18–21], $M_{\mathrm{shock}} \lesssim 10^{12} M_{\odot}$. Such haloes are encompassed by a stable shock near their outer radius, R_{v}, inside which gravity and thermal energy are in virial equilibrium. Gas falling in through the shock is expected to heat up to the virial temperature and stall in quasi-static equilibrium before it cools and descends into the inner galaxy[22]. However, at $z \geq 2$, these hot haloes are penetrated by cold streams[8,9,20]. Because early haloes with $M_{\mathrm{v}} > M_{\mathrm{shock}}$ populate the massive tail of the distribution, they are fed by dark matter filaments from the cosmic web that are narrow in comparison with R_{v} and denser than the mean within the halo[8]. The enhanced density of the gas along these filaments makes the flows along them unstoppable; in particular, they cool before they develop the pressure to support a shock, and thus avoid shock heating (Supplementary Information, section 2).

To investigate the penetration of cold streams, we study the way gas feeds massive high-z galaxies in the cosmological MareNostrum simulation—an adaptive-mesh hydrodynamical simulation in a co-moving box of side length 71 Mpc and a resolution of 1.4 kpc at the galaxy centres (Supplementary Information, section 3). The gas maps in Figs 1 and 2 demonstrate how the shock-heated, high-entropy, low-flux medium that fills most of the halo is penetrated by three narrow, high-flux streams of low-entropy gas (Supplementary Figs 3–6). The penetration is evaluated from the profiles of gas inflow rate, $\dot{M}(r)$, through shells of radius r (Fig. 3, Supplementary Fig. 7). The average profile reveals that the flow rate remains constant from well outside $R_{\mathrm{v}} \approx 90$ kpc to the disk inside $r \approx 15$ kpc.

To relate the feeding by streams to the observed abundance of galaxies as a function of SFR, we use the MareNostrum inflow-rate profiles to evaluate $n(>\dot{M})$, the co-moving number density of galaxies with an inflow rate $>\dot{M}$. We first extract the conditional probability distribution $P(\dot{M} \mid M_{\mathrm{v}})$ by sampling the $\dot{M}(r)$ profiles

[1]Racah Institute of Physics, The Hebrew University, Jerusalem 91904, Israel. [2]Harvard Smithsonian Center for Astrophysics, 60 Garden St, Cambridge, Massachusetts 02138, USA. [3]Departement de Physique, ENS, 24 rue Lhomond, 75231 Paris cedex 05, France. [4]Max Planck Institute for Astrophysics, Karl-Schwarzschild-Strasse 1, 85741 Garching, Germany. [5]Institut d'Astrophysique de Paris and UPMC, 98bis Boulevard Arago, Paris 75014, France. [6]CEA Saclay, DSM/IRFU, UMR AIM, Batiment 709, 91191 Gif-sur-Yvette cedex, France. [7]Institute for Theoretical Physics, University of Zurich, CH-8057 Zurich, Switzerland.

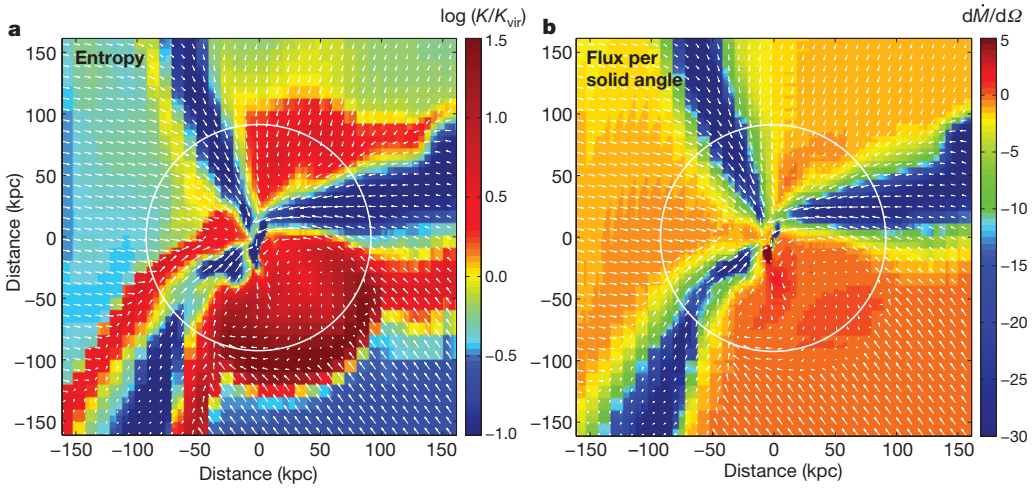


Figure 1 | Entropy, velocity and inward flux of cold streams penetrating hot haloes. **a**, **b**, Maps referring to a thin slice through one of our fiducial galaxies with $M_v = 10^{12} M_\odot$ at $z = 2.5$. The arrows describe the velocity field, scaled such that the distance between the tails is 260 km s^{-1}. The circle marks the halo virial radius, R_v. The entropy, $\log K = \log(T/\rho^{2/3})$, in units of the virial quantities, highlights (in red) the high-entropy medium filling the halo out to the virial shock outside R_v. It exhibits (in blue) three radial, low-entropy streams that penetrate the inner disk, seen edge-on. The radial flux per solid angle is $\dot{m} = r^2 \rho v_r$, in solar masses per year per square radian, where ρ is the gas density and v_r the radial velocity. It demonstrates that more than 90% of the inflow is channelled through the streams (blue), at a rate that remains roughly the same at all radii. This rate is several times higher than the spherical average outside the virial sphere, $\dot{m}_{vir} \approx 8 M_\odot\,\mathrm{yr}^{-1}\,\mathrm{rad}^{-2}$, according to equation (1). The opening angle of a typical stream at R_v is $20^\circ - 30^\circ$, so the streams cover a total angular area of ~ 0.4 rad^2, namely a few per cent of the sphere. When viewed from a given direction, the column density of cold gas below 10^5 K is above 10^{20} cm^{-2} for 25% of the area within the virial radius. Although the pictures show the inner disk, the disk width is not resolved, so associated phenomena such as shocks, star formation and feedback are treated in an approximate way only (see density maps and additional cases in Supplementary Figs 3–5). K_{vir}, virial entropy.

uniformly in r, using the fact that the velocity along the streams is roughly constant (Supplementary Information, sections 5 and 6). This is convolved with the halo mass function[23], $n(M_v)$, to give

$$n(\dot{M}) = \int_0^\infty P(\dot{M}\,|M_v)\, n(M_v)\, \mathrm{d}M_v$$

The desired cumulative abundance, $n(>\dot{M})$, obtained by integration over the inflow rates from $\dot{M}$ to infinity, is shown at $z = 2.2$ in Fig. 4. Assuming that the SFR equals $\dot{M}$, the curve referring to $\dot{M}$ lies safely above the observed values, marked by the symbols, indicating that the gas input rate is sufficient to explain the SFR. However, $\dot{M}$ and the SFR are allowed to differ only by a factor of ~ 2, confirming our suspicion that the SFR must closely follow the gas input rate. The simulated SFR indeed traces the accretion rate to within a factor of two, but, given that our disks are poorly resolved, we focus here on the accretion as the more robustly simulated quantity. Because at $z \approx 2.2$ the star-forming galaxies constitute only a fraction of the observed $\sim 10^{11} M_\odot$ galaxies[24,25], the requirement for a SFR almost as great as $\dot{M}$, based on Fig. 4, becomes even stronger.

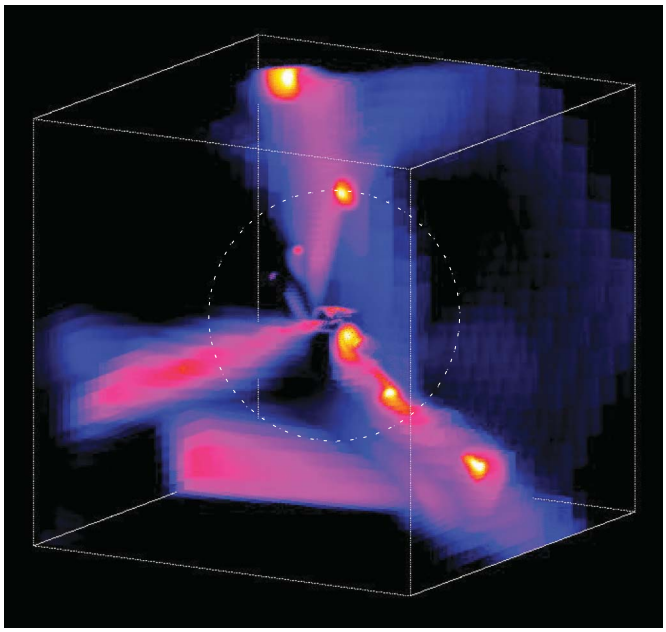


Figure 2 | Streams in three dimensions. The map shows radial flux for the galaxy of Fig. 1 in a box of side length 320 kpc. The colours refer to inflow rate per solid angle of point-like tracers at the centres of cubic-grid cells. The dotted circle marks the halo virial radius. The appearance of three fairly radial streams seems to be generic in massive haloes at high redshift, and is a feature of the cosmic web that deserves an explanation. Two of the streams show gas clumps of mass on the order of one-tenth that of the central galaxy, but most of the stream mass is smoother (Supplementary Fig. 6). The $\gtrsim 10^{10} M_\odot$ clumps, which involve about one-third of the incoming mass, are also gas rich—in the current simulation only 30% of their baryons turn into stars before they merge with the central galaxy.

By analysing the clumpiness of the gas streams, using the sharp peaks of inflow in the $\dot{M}(r)$ profiles, we address the role of mergers versus smooth flows. We evaluate each clump mass by integrating $M_{clump} = \int (\dot{M}(r)/v_r(r))\, \mathrm{d}r$ across the peak, and estimate a mass ratio for the expected merger as $\mu = M_{clump}/f_b M_v$, ignoring further mass loss in the clump on its way in and deviations of the galaxy baryon fraction from f_b. We use 'merger' to describe any major or minor merger with $\mu \geq 0.1$, as distinct from 'smooth' flows, which include 'mini-minor' mergers with $\mu < 0.1$. We find that about one-third of the mass is flowing in as mergers and the rest as smoother flows. However, the central galaxy is fed by a clump with $\mu \geq 0.1$ less than 10% of the time; that is, the duty cycle for mergers is $\eta \lesssim 0.1$. A similar estimate is obtained using EPS merger rates[7] and starburst durations of ~ 50 Myr at $z = 2.5$ from simulations[26] (Supplementary Information, section 5).

From the difference between the two curves of Fig. 4, we learn that only one-quarter of the galaxies with a given $\dot{M}$ are to be seen during a merger. The fact that the SFGs lie well above the merger curve even if the SFR is $\sim \dot{M}$ indicates that in most of them the star formation is driven by smooth streams. Thus, 'SFG' could also stand for 'stream-fed galaxy'. This may explain why these galaxies maintain an

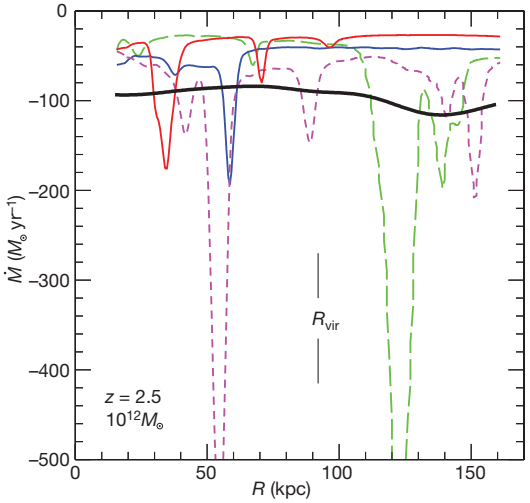


Figure 3 | Accretion profiles, $\dot{M}(r)$. Shown is the gas inflow rate through spherical shells of radius r, from the disk vicinity to almost twice the halo virial radius, obtained by integrating $r^2\rho v_r$ over the whole shell. The thick black curve is the average over the simulated galaxies of the fiducial case, $M_v \approx 10^{12}M_\odot$ at $z = 2.5$. It shows deep penetration at a roughly constant rate of $\sim 100M_\odot$ yr^{-1}, consistent with the virial growth rate predicted by equation (1). Apparently, the inflow rate does decay as the gas travels through the halo, but this decay is roughly compensated for by the higher cosmological inflow rate when that gas entered the halo (equation (1)), leading to the apparent constancy of accretion rate with radius. The coloured curves refer to four representative galaxies, two showing clumps with $\mu \gtrsim 0.1$ (dashed lines) and two with smoother flows involving only mini-minor clumps with $\mu < 0.1$ (solid lines). Clumps with $\mu \gtrsim 0.3$ appear within $2R_v$ about once in every ten galaxies; that is, major mergers are infrequent (Supplementary Fig. 7). The $\dot{M}(r)$ profiles serve for extracting the conditional probability distribution $P(\dot{M} \mid M_v)$, leading to the abundance $n(>\dot{M})$ (Supplementary Fig. 8).

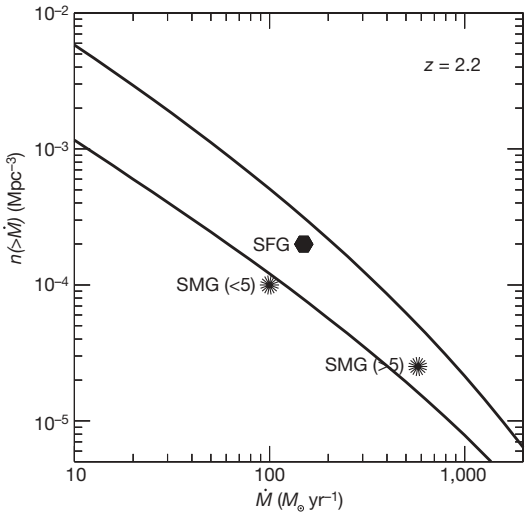


Figure 4 | Abundance of galaxies as a function of gas inflow rate, $n(>\dot{M})$. Shown is the co-moving number density, n, of galaxies with inflow rate higher than $\dot{M}$ at $z = 2.2$, as predicted from our analysis of the cosmological simulation. The upper curve refers to total inflow. It shows that galaxies with $\dot{M} > 150M_\odot$ yr^{-1} are expected at a co-moving number density $n \approx 3 \times 10^{-4}$ Mpc^{-3} (similar to estimates in other simulations[32,33]). Fluxes as high as $\dot{M} > 500M_\odot$ yr^{-1} are anticipated at $n \approx 6 \times 10^{-5}$ Mpc^{-3}. The lower curve is similar, but limited to gas input by $\mu > 0.1$ mergers. The symbols represent the vicinity of where the observed massive star-forming galaxies can be located once their observed SFRs are identified with $\dot{M}$. The sBzK and BX/BM galaxies are marked SFG[13]. The SMGs respectively brighter and fainter than 5 mJy are marked accordingly[12,13]. We see that the overall gas inflow rate is sufficient for the observed SFR, but the small margin implies that the SFR must closely follow the rate of gas supply. Most of the massive star formers at a given SFR are expected to be observed while being fed by smooth flows rather than undergoing mergers. By studying the contribution of different halo masses to the abundance $n(>\dot{M})$, we learn that the high-SFR SFGs and SMGs are associated with haloes of mass $10^{12}M_\odot - 10^{13}M_\odot$ (Supplementary Fig. 9). An integration of $\dot{M}$ over halo mass and time reveals that most of the stars in the universe were formed in stream-fed galaxies, within haloes of mass $>2 \times 10^{11}M_\odot$ at $1.5 < z < 4$.

extended, thick disk while doubling their mass over a halo crossing time[4]. On the other hand, if the SFR is $\sim\dot{M}$, we learn from Fig. 4 that about half of the bright SMGs and most of the fainter SMGs lie below the merger curve and are therefore consistent with being merger-induced starbursts[13].

We find that stream-fed galaxies of mass $\sim 10^{11}M_\odot$ at $z \approx 2.5$ were the most productive star formers in the universe. The constraints on the overall SFR density at these epochs imply that SFR has been suppressed in smaller galaxies, for example by photo-ionization and stellar feedback[8,27,28]. The early presence of low-SFR galaxies[24,25] requires quenching of SFR also at the massive end, perhaps due to gravitational heating by destructive streams[29].

The streams are likely also to be responsible for compact spheroids, as an alternative to mergers and the associated heating by expanding shocks[22,29]. Using equation (1), we find that at $z \geq 2$ the streams can maintain both the high gas fraction and the turbulence necessary for the disk to break up into giant clumps by gravitational instability, with dispersion-to-rotation ratio $\sigma/V \approx 0.25$, as observed[4,30]. The clumps migrate inward and dissipatively merge into a spheroid[10,11]. The stream carrying the largest coherent flux with an impact parameter of a few kiloparsecs determines the disk's spin and orientation, and the stream clumps perturb it. The incoming clumps and the growing spheroid can eventually stabilize the disk and suppress star formation. We can thus associate the streams with the main mode of galaxy and star formation occurring in massive haloes at $z \approx 2-3$; the streams that create the disks also make them fragment into giant clumps that serve both as the sites of efficient star formation and the progenitors of the central spheroid, which in turn helps the streams to quench star formation.

Although wet mergers may grow secondary disks[31], they are not as frequent as the observed SFGs (Fig. 4), these disks are neither gas rich nor clumpy enough, and, unlike most SFGs, they are dominated by stellar spheroids.

The cold streams should be detectable by absorption or emission. For external background sources, our simulation predicts that haloes with $M_v \approx 10^{12}M_\odot$ at $z \approx 2.5$ should contain gas at temperature $<10^5$ K with column densities $>10^{20}$ cm^{-2} covering $\sim$25% of the area at radii between 20 and 100 kpc, with coherent velocities of $\lesssim 200$ km s^{-1}. Sources at the central galaxies should show absorption by the radial streams in $\sim$5% of the galaxies, flowing in at $\gtrsim 200$ km s^{-1}, with column densities $\sim 10^{21}$ cm^{-2} (Supplementary Figs 3–5).

Received 30 July; accepted 7 November 2008.

1. Genzel, R. *et al.* The rapid formation of a large rotating disk galaxy three billion years after the Big Bang. *Nature* **442**, 786–789 (2006).
2. Chapman, S. C., Smail, I., Blain, A. W. & Ivison, R. J. A population of hot, dusty ultraluminous galaxies at $z \sim 2$. *Astrophys. J.* **614**, 671–678 (2004).
3. Förster Schreiber, N. M. *et al.* SINFONI integral field spectroscopy of $z \sim 2$ UV-selected galaxies: Rotation curves and dynamical evolution. *Astrophys. J.* **645**, 1062–1075 (2006).
4. Genzel, R. *et al.* From rings to bulges: evidence for rapid secular galaxy evolution at $z \sim 2$ from integral field spectroscopy in the SINS survey. *Astrophys. J.* **687**, 59–77 (2008).
5. Stark, D. P. *et al.* The formation and assembly of a typical star-forming galaxy at $z \approx 3$. *Nature* **455**, 775–777 (2008).
6. Neistein, E., van den Bosch, F. C. & Dekel, A. Natural downsizing in hierarchical galaxy formation. *Mon. Not. R. Astron. Soc.* **372**, 933–948 (2006).
7. Neistein, E. & Dekel, A. Merger rates of dark-matter haloes. *Mon. Not. R. Astron. Soc.* **388**, 1792–1802 (2008).
8. Dekel, A. & Birnboim, Y. Galaxy bimodality due to cold flows and shock heating. *Mon. Not. R. Astron. Soc.* **368**, 2–20 (2006).

9. Kereš, D., Katz, N., Weinberg, D. H. & Davé, R. How do galaxies get their gas? *Mon. Not. R. Astron. Soc.* **363,** 2–28 (2005).
10. Noguchi, M. Early evolution of disk galaxies: Formation of bulges in clumpy young galactic disks. *Astrophys. J.* **514,** 77–95 (1999).
11. Elmegreen, B., Bournaud, F. & Elmegreen, D. M. Bulge formation by the coalescence of giant clumps in primordial disk galaxies. *Astrophys. J.* **688,** 67–77 (2008).
12. Wall, J. V., Pope, A. & Scott, D. The evolution of submillimetre galaxies: two populations and a redshift cut-off. *Mon. Not. R. Astron. Soc.* **383,** 435–444 (2008).
13. Tacconi, L. J. *et al.* Submillimeter galaxies at $z \sim 2$: Evidence for major mergers and constraints on lifetimes, IMF, and CO-H_2 conversion factor. *Astrophys. J.* **680,** 246–262 (2008).
14. Adelberger, K. L. *et al.* Optical selection of star-forming galaxies at redshifts $1 < z$ 3. *Astrophys. J.* **607,** 226–240 (2004).
15. Daddi, E. *et al.* A new photometric technique for the joint selection of star-forming and passive galaxies at $1.4 < z$ 2.5. *Astrophys. J.* **617,** 746–764 (2004).
16. Neistein, E. & Dekel, A. Constructing merger trees that mimic N-body simulations. *Mon. Not. R. Astron. Soc.* **383,** 615–626 (2008).
17. Genel, S. *et al.* Mergers and mass accretion rates in galaxy assembly: The millennium simulation compared to observations of $z \sim 2$ galaxies. *Astrophys. J.* **688,** 789–793 (2008).
18. Birnboim, Y. & Dekel, A. Virial shocks in galactic haloes? *Mon. Not. R. Astron. Soc.* **345,** 349–364 (2003).
19. Binney, J. On the origin of the galaxy luminosity function. *Mon. Not. R. Astron. Soc.* **347,** 1093–1096 (2004).
20. Ocvirk, P., Pichon, C. & Teyssier, R. Bimodal gas accretion in the MareNostrum galaxy formation simulation. *Mon. Not. R. Astron. Soc.* **390,** 1326–1338 (2008).
21. Keres, D. *et al.* Galaxies in a simulated ΛCDM Universe I: cold mode and hot cores. Preprint at ⟨http://arxiv.org/abs/0809.1430⟩ (2008).
22. Birnboim, Y., Dekel, A. & Neistein, E. Bursting and quenching in massive galaxies without major mergers or AGNs. *Mon. Not. R. Astron. Soc.* **380,** 339–352 (2007).
23. Sheth, R. K. & Tormen, G. An excursion set model of hierarchical clustering: ellipsoidal collapse and the moving barrier. *Mon. Not. R. Astron. Soc.* **329,** 61–75 (2002).
24. Kriek, M. *et al.* Spectroscopic identification of massive galaxies at $z \sim 2.3$ with strongly suppressed star formation. *Astrophys. J.* **649,** L71–L74 (2006).
25. van Dokkum, P. G. *et al.* Confirmation of the remarkable compactness of massive quiescent galaxies at $z \sim 2.3$: Early-type galaxies did not form in a simple monolithic collapse. *Astrophys. J.* **677,** L5–L8 (2008).
26. Cox, T. J., Jonsson, P., Somerville, R. S., Primack, J. R. & Dekel, A. The effect of galaxy mass ratio on merger-driven starbursts. *Mon. Not. R. Astron. Soc.* **384,** 386–409 (2008).
27. Dekel, A. & Silk, J. The origin of dwarf galaxies, cold dark matter, and biased galaxy formation. *Astrophys. J.* **303,** 39–55 (1986).
28. Dekel, A. & Woo, J. Feedback and the fundamental line of low-luminosity low-surface-brightness/dwarf galaxies. *Mon. Not. R. Astron. Soc.* **344,** 1131–1144 (2003).
29. Dekel, A. & Birnboim, Y. Gravitational quenching in massive galaxies and clusters by clumpy accretion. *Mon. Not. R. Astron. Soc.* **383,** 119–138 (2008).
30. Elmegreen, D. M., Elmegreen, B. G. & Hirst, A. C. Discovery of face-on counterparts of chain galaxies in the Tadpole Advanced Camera for Surveys Field. *Astrophys. J.* **604,** L21–L23 (2004).
31. Robertson, B. E. & Bullock, J. S. High-redshift galaxy kinematics: Constraints on models of disk formation. *Astrophys. J.* **685,** L27–L30 (2004).
32. Finlator, K., Davé, R., Papovich, C. & Hernquist, L. The physical and photometric properties of high-redshift galaxies in cosmological hydrodynamic simulations. *Astrophys. J.* **639,** 672–694 (2006).
33. Nagamine, K., Ouchi, M., Springel, V. & Hernquist, L. Lyman-alpha emitters and Lyman-break galaxies at $z = 3$–6 in cosmological SPH simulations. Preprint at ⟨http://arxiv.org/abs/0802.0228⟩ (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We acknowledge discussions with N. Bouche, S. M. Faber, R. Genzel, D. Koo, A. Kravtsov, A. Pope, J. R. Primack, J. Prochaska, A. Sternberg and J. Wall. This research was supported by the France–Israel Teamwork in Sciences, the German–Israel Science Foundation, the Israel Science Foundation, a NASA Theory Program at UCSC, and a Minerva fellowship (T.G.). We thank the Barcelona Centro Nacional de Supercomputación for computer resources and technical support. The simulation is part of the Horizon collaboration.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to A.D. (dekel@phys.huji.ac.il).

LETTERS

High-*Q* surface-plasmon-polariton whispering-gallery microcavity

Bumki Min[1,2]†, Eric Ostby[1], Volker Sorger[2], Erick Ulin-Avila[2], Lan Yang[1]†, Xiang Zhang[2,3] & Kerry Vahala[1]

Surface plasmon polaritons (SPPs) are electron density waves excited at the interfaces between metals and dielectric materials[1]. Owing to their highly localized electromagnetic fields, they may be used for the transport and manipulation of photons on subwavelength scales[2–9]. In particular, plasmonic resonant cavities represent an application that could exploit this field compression to create ultra-small-mode-volume devices. A key figure of merit in this regard is the ratio of cavity quality factor, *Q* (related to the dissipation rate of photons confined to the cavity), to cavity mode volume, *V* (refs 10, 11). However, plasmonic cavity *Q* factors have so far been limited to values less than 100 both for visible and near-infrared wavelengths[12–16]. Significantly, such values are far below the theoretically achievable *Q* factors for plasmonic resonant structures. Here we demonstrate a high-*Q* SPP whispering-gallery microcavity that is made by coating the surface of a high-*Q* silica microresonator with a thin layer of a noble metal. Using this structure, *Q* factors of 1,376 ± 65 can be achieved in the near infrared for surface-plasmonic whispering-gallery modes at room temperature. This nearly ideal value, which is close to the theoretical metal-loss-limited *Q* factor, is attributed to the suppression and minimization of radiation and scattering losses that are made possible by the geometrical structure and the fabrication method. The SPP eigenmodes, as well as the dielectric eigenmodes, are confined within the whispering-gallery microcavity and accessed evanescently using a single strand of low-loss, tapered optical waveguide[17,18]. This coupling scheme provides a convenient way of selectively exciting and probing confined SPP eigenmodes. Up to 49.7 per cent of input power is coupled by phase-matching control between the microcavity SPP and the tapered fibre eigenmodes.

The subject of high-*Q* optical micro- and nanocavities has been intensively investigated over the last decade. The extremely low photon loss rate and small cavity mode volume of photonic-crystal or whispering-gallery devices offer surprisingly rich physics, spanning many areas of research including nonlinear optics, quantum optics, and device physics[10,11]. Whereas optical micro- and nanocavities made of dielectric or semiconducting materials exhibit large *Q* factors as well as small diffraction-limited cavity mode volumes, their metallic counterparts (surface-plasmonic cavities[12–16]) have been optimized primarily for subwavelength-scale miniaturization and have given results well below the theoretically predicted performance limit—especially in terms of cavity loss—set by ohmic loss in the metal. This is believed to result from other loss contributions such as surface scattering, radiation, finite cavity mirror reflectance or a significant degree of field penetration into the metal. However, these seemingly distinct dielectric and plasmonic waveguiding principles can be combined in a single cavity by using mature optical microcavity technology such as that provided by disk[19,20] or toroidal microcavities[21]. Here we propose to utilize a dielectric microcavity, engineered to minimize surface blemishes and thereby reduce scattering[19,20], as a template for the creation of a surface-plasmonic whispering-gallery microcavity with a cavity plasmon-polariton loss rate close to the theoretical limit.

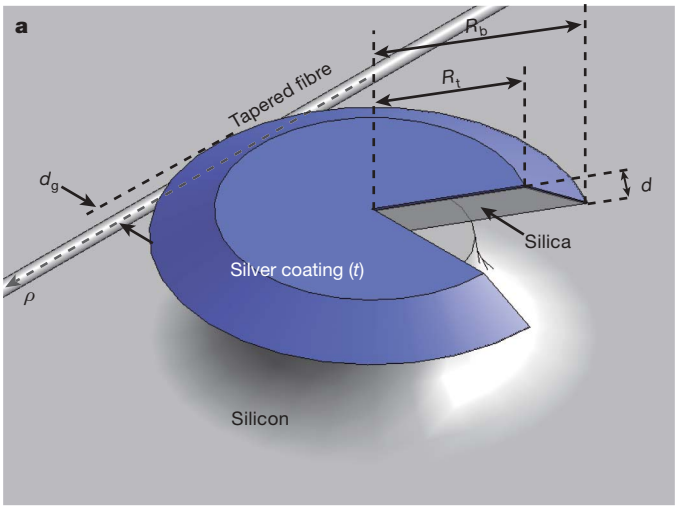
The proposed plasmonic microdisk cavity structure is sketched in Fig. 1a. The plasmonic cavity is composed of a core silica (silicon dioxide) disk microcavity clad in a thin layer of silver (see Methods). Silica microdisk resonators are ideal templates for the study of surface-plasmonic whispering-gallery modes primarily because they routinely have optical *Q* factors greater than 1,000,000 (considerably larger than the metal-loss-limited *Q* factor). Using the wedge structure shown in Fig. 1a, *Q* factors as high as 6×10^7 have been demonstrated, showing a remarkably low scattering loss value[19]. A scanning electron micrograph of a silver-coated SPP microdisk resonator is shown in Fig. 1b, and the corresponding expanded view of the edge of the disk resonator is shown in Fig. 1c.

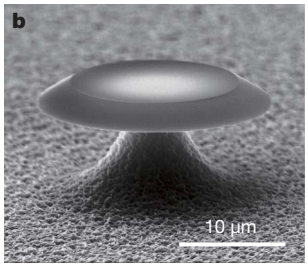
A full vectorial finite-element analysis was performed for the SPP microdisk resonators[22,23], taking into account the effects of silver[24] and silica[25] material dispersion. The theoretical cavity mode dispersion diagram of an SPP microdisk resonator (Fig. 2a) shows the real part of the eigenfrequency, f, of the cavity modes as a function of an azimuthal mode number, m. The vacuum light line is defined by $f = mc/2\pi R_b$ with respect to the bottom radius, R_b, of the template silica disk microcavity, and the silica light line is similarly defined by $f = mc/2\pi n_{silica}(f)R_b$. Here c is the speed of light and n_{silica} is the refractive index of silica. The eigenmodes of an SPP microcavity can be classified into two distinctive categories in terms of the cavity mode dispersion: (1) surface-plasmonic modes at the metal–dielectric interface and (2) optical dielectric modes due to the presence of a dielectric waveguiding channel.

In the insets of Fig. 2a, the fundamental (first-order) SPP eigenmode, the second-order SPP eigenmode and the fundamental dielectric eigenmode are plotted for magnetic energy density $u_M = (1/2\mu_0)|\mathbf{B}(r, \phi, z)|^2$ (where μ_0 is the permeability of free space) using a false-colour map (a conventional cylindrical coordinate system (r, ϕ, z) is used for the analysis). The SPP eigenmodes of an SPP microdisk resonator have electromagnetic energy-density profiles that peak at the silica–metal interface in the transverse plane (constant ϕ). The SPP eigenmodes are categorized as SPP_{qm}, where q is the plasmonic mode number ($\mathbf{H}(r, \phi, z) = \mathbf{H}^{qm}_{\mathrm{SPP}}(r, z)e^{im\phi}$), and the optical dielectric eigenmodes are denoted by DE_{hm}, where h is the dielectric mode number ($\mathbf{H}(r, \phi, z) = \mathbf{H}^{hm}_{\mathrm{DE}}(r, z)e^{im\phi}$). The plasmonic mode number is defined as the number of antinodes in $|\mathbf{H}^{qm}_{\mathrm{SPP}}|$ along the silica–metal interface (excluding the vicinity of the sharp corner of the microcavity). Dispersion relations for the four lowest-order SPP eigenmodes ($q = 1, 2, 3, 4$) and the two lowest-order dielectric eigenmodes ($h = 1, 2$; see Methods) are plotted in Fig. 2a.

The cavity mode index, n_c, of a specific eigenmode can be evaluated with respect to the dielectric cavity edge ($r = R_b$) as $n_c = mc/2\pi R_b f$.

[1]Thomas J. Watson Laboratory of Applied Physics, California Institute of Technology, Pasadena, California 91125, USA. [2]Nanoscale Science and Engineering Center, 5130 Etcheverry Hall, University of California, Berkeley, California 94720, USA. [3]Material Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. †Present addresses: Department of Mechanical Engineering, Korea Advanced Institute of Science and Technology, Daejeon 305-751, Republic of Korea (B.M.); Department of Electrical and Systems Engineering, Washington University in St Louis, St Louis, Missouri 63130, USA (L.Y.).





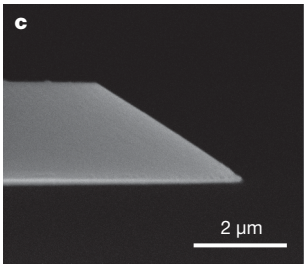
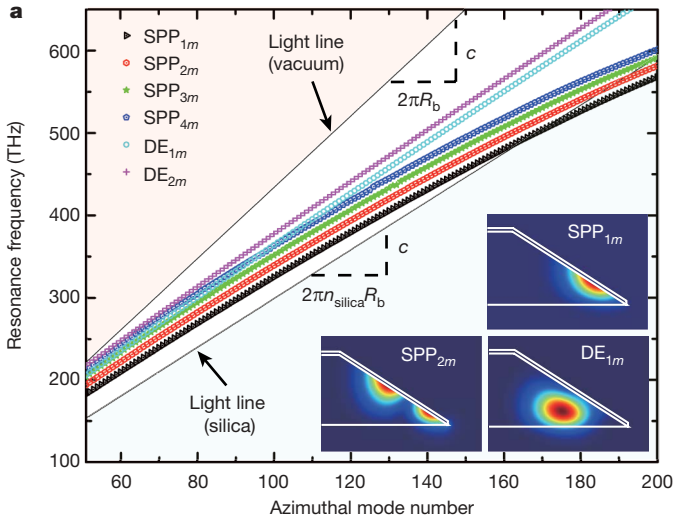
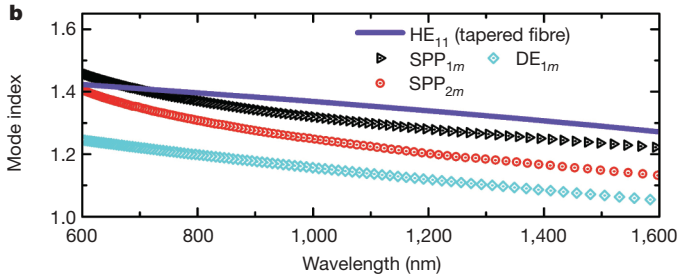


Figure 1 | Tapered fibre waveguide and SPP whispering-gallery microdisk resonator. a, SPP microdisk resonator with a tapered optical fibre passing under its edge. The wedge-shaped disk edge is a by-product of isotropic buffered hydrofluoric acid etching of silica. A transverse cross-section of the cavity is shown for clarity. R_b, bottom radius; R_t, top radius; d, thickness of the silica disk resonator; t, thickness of the metal layer. The straight fibre waveguide axis is denoted by the coordinate ρ and the gap width, d_g, is defined as the horizontal distance from the dielectric cavity edge to the fibre axis. **b**, Scanning electron micrograph of a fabricated silver-coated SPP microdisk resonator ($R_b = 10.96$ μm, $R_t = 7.89$ μm, $d = 2$ μm, $t \approx 100$ nm). **c**, Expanded view of the edge of the SPP microdisk resonator.





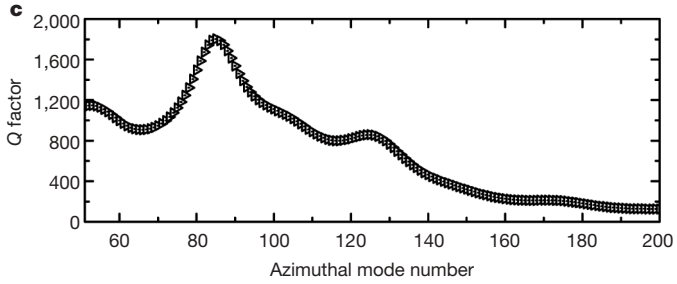


Figure 2 | Cavity mode dispersion, effective mode index and *Q* factor. a, Cavity mode dispersion curves for an SPP microdisk resonator, calculated from finite-element eigenfrequency analysis. For this calculation, the thickness of the silver layer is 100 nm, and the bottom and top radii and the thickness of the template silica microdisk resonator were set to 11, 7.9 and 2 μm, respectively. Light lines, corresponding to vacuum and silica, are given as two black lines (silica material dispersion has been taken into account). For clarity, only the four lowest-order SPP eigenmodes and the two lowest-order dielectric eigenmodes are plotted. The first- and second-order SPP eigenmodes (SPP_{1m}, SPP_{2m}) and the fundamental dielectric eigenmode (DE_{1m}) are shown in the inset. **b**, Effective cavity mode indices, n_c, of SPP_{1m}, SPP_{2m} and DE_{1m} (with respect to R_b), shown as a function of resonance wavelength. The mode index of a tapered-fibre HE_{11} mode is shown to demonstrate phase matching. **c**, The theoretical *Q* factor for SPP_{1m}, plotted as a function of azimuthal mode number, m.

Figure 2b shows the calculated mode index for modes SPP_{1m}, SPP_{2m} and DE_{1m}. The mode index of a fundamental surface-plasmonic mode (SPP_{1m}) is clearly larger than that of a fundamental dielectric mode (DE_{1m}) within most of the visible and near-infrared frequency band, owing to the plasmonic surface-wave characteristics. The mode index is important because it determines the phase-matching condition for excitation of SPP modes by an input tapered fibre waveguide. After n_c has been calculated, the corresponding phase-matched fibre-waveguide mode index can be approximated as

$$n_w \approx n_c \frac{\sin^{-1}\sqrt{\delta(2-\delta)}}{\sqrt{\delta(2-\delta)}} = n_c\left(1+\frac{1}{3}\delta+\frac{2}{15}\delta^2+O(\delta^3)\right) \quad (1)$$

where $\delta = -d_g/R_b \geq 0$ denotes the relative gap width (d_g, gap width; see Methods). To qualitatively describe the effect of gap width variation on the phase matching, the HE_{11} mode index of a fibre waveguide with a 1-μm waist diameter is shown in Fig. 2b. The fibre mode index is slightly larger than the SPP_{1m} mode index in the near-infrared wavelength band. However, owing to the above phase-matching formula, the SPP_{1m} eigenmode can be effectively phase-matched to the tapered-fibre eigenmode by increasing the relative gap width. We also note that the diameter of the tapered fibre can be optimized to phase-match the cavity eigenmodes to the fibre eigenmode because the fibre diameter determines the mode index of the fibre eigenmode.

The calculated cavity *Q* factors for SPP_{1m} eigenmodes as a function of azimuthal mode number, m, are plotted in Fig. 2c. The error bounds for the imaginary part of the permittivity of silver are taken into account in estimating the bounds on the theoretical *Q* factors and are discussed in Supplementary Information[24]. The calculated *Q* factors consist of contributions from intrinsic metal loss (silica material loss is negligible in comparison with metal loss[22,24,25]) and the geometry- and material-dependent radiation loss into free space: $Q^{-1} \approx Q_{metal}^{-1} + Q_{rad}^{-1}$. Therefore, this *Q* value provides the ideal theoretical limit on the *Q* performance of SPP microdisk resonators that have negligible scattering loss induced by surface roughness. The radiation-limited *Q* factor, Q_{rad}, is orders of magnitude larger than the metal-loss-limited *Q* factor, Q_{metal}; the ideal SPP microcavity is thus metal-loss limited (see Methods): $Q^{-1} \approx Q_{metal}^{-1}$. In Fig. 2c, the highest fundamental SPP *Q* factor is found to be 1,800, at the resonant wavelength of 1,062.45 nm ($m = 85$). At a

wavelength of 1,568.25 nm ($m = 54$), which is close to the value used in measurements described below (series 1 in Fig. 3), the theoretical Q factor is 1,140 (see Supplementary Information for the lower and upper bounds, $Q_l = 700$ and $Q_u = 2{,}210$). The cavity mode volume, V, and the figure of merit Q/V of the SPP microcavity are estimated in Supplementary Information.

To measure the SPP microdisk resonances experimentally, a narrow-linewidth (<300 kHz) tunable external-cavity semiconductor laser is coupled to the tapered fibre waveguide and scanned over the 1,520–1,570-nm wavelength range. The position of the tapered fibre with respect to the SPP microdisk resonator is controlled at a fixed vertical distance by piezoelectric stages with an encoded resolution of 100 nm, and the laser polarization is controlled using a fibre polarization controller and monitored with a polarimeter. For large overlap between the cavity and the waveguide modes, the tapered fibre is positioned underneath the bevelled edge of the resonator, where the silica microdisk is free of silver coating. The output transmission is recorded using a photodetector and a digital oscilloscope. Figure 3a shows the normalized transmission spectrum from an SPP microdisk resonator with a Lorentzian line-shape fit (Fig. 3a, red curve) to each resonance. Two resonances, located at 1,523.59 and 1,532.76 nm ($SPP_{1,83}$ and $DE_{1,74}$, as estimated by calculation), can be clearly identified. An expanded view of the scan (main panel modes outlined) is shown in the inset of Fig. 3a

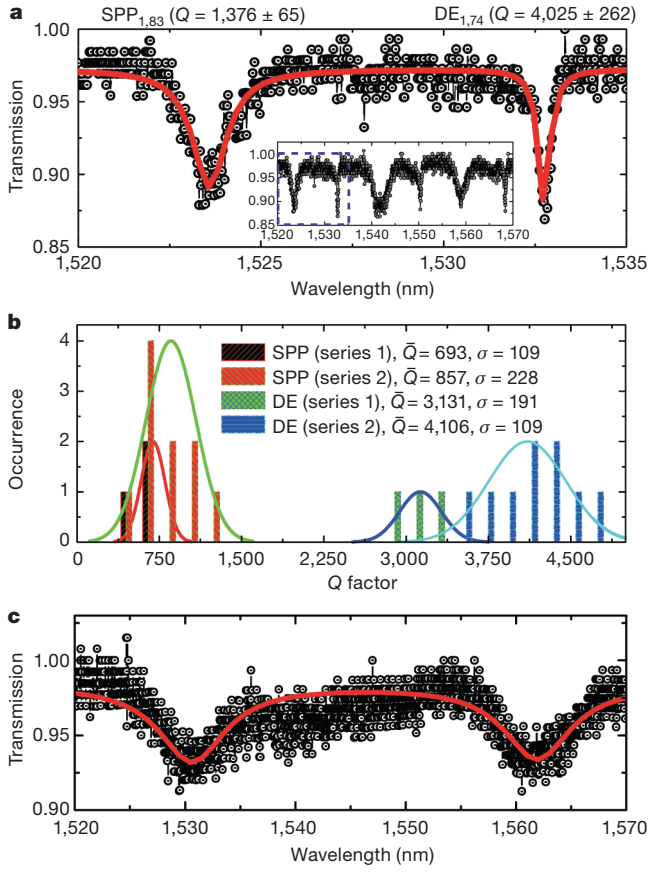


Figure 3 | Q-factor measurements for silver-coated and chromium-coated microdisk resonators. **a**, Normalized transmission spectrum showing the highest measured SPP Q factor of 1,376 ± 65 and a dielectric resonance with a Q factor of 4,025 ± 262. The inset (main panel outlined) is the entire wavelength band (1,520–1,570 nm) scanned for this sample. $R_b = 15.45 \pm 0.05$ μm, $R_t = 12.73 \pm 0.04$ μm, $d = 2$ μm, $t \approx 100$ nm. **b**, Statistical histogram of measured Q values showing the occurrence of each eigenmode (SPP and dielectric) for two different sample batches (series 1 and series 2). Mean ($\bar{Q}$) and standard deviation (σ) of Q factors are shown in the key (series 1, $n = 3$ measurements; series 2, $n = 9$). **c**, Normalized transmission spectrum for a chromium-coated microdisk resonator ($R_b \approx 11$ μm, $R_t \approx 7.9$ μm, $d = 2$ μm) with a two-dip Lorentzian fit.

and spans three free spectral ranges of SPP and dielectric eigenmodes. The cavity Q factor for the fundamental $SPP_{1,83}$ eigenmode is found to be 1,376 ± 65 (which falls within the theoretical Q-factor range of $760 \lesssim Q \lesssim 2{,}360$, with a nominal Q factor of 1,225 for the $SPP_{1,83}$ eigenmode), and that of the fundamental $DE_{1,74}$ mode is 4,025 ± 262. This SPP Q factor of ~1,376 is over 30 times larger than the Q factors reported in previous SPP cavity work[12–16].

To determine the reproducibility of this Q factor, two series of samples of different nominal sizes (series 1, $R_b = 10.93$ μm; series 2, $R_b = 15.56$ μm) were tested. The measured Q factors for both the SPP and dielectric eigenmodes in the 1,550-nm wavelength band are plotted statistically in Fig. 3b. Two separate clusters of Q factors are seen in this plot, indicating the distinctive resonant characteristics of the two sorts of eigenmode and a tendency for loss to decrease (Q factor to increase) as the size of the cavity increases. The measured Q factors are bounded within the range predicted for the plasmonic eigenmode (see Supplementary Information). To test the metal-dependent resonance characteristics of the SPP microdisk, chromium (which is highly lossy at optical frequencies) was deposited onto the silica microdisk using the same sputtering process, for use in control experiments. The normalized transmission spectrum for a chromium-coated microdisk resonator is shown in Fig. 3c. In this case, only low-Q resonances (for example $Q \approx 213$ at 1,561.92 nm) are observed, owing to the presence of the chromium layer. These resonances are primarily of optical dielectric origin, as confirmed by finite-element simulations, because the fundamental SPP eigenmodes of a chromium-coated microdisk of this size should have a theoretical Q factor of ~10 in the 1,550-nm band.

To verify the phase-matched excitation of the cavity eigenmodes, a series of measurements were performed with variations in the position of the tapered fibre waveguide relative to the SPP cavity. Figure 4 shows the normalized transmission spectra (for an SPP microdisk from a batch from series 2) excited at different gap widths, d_g, and also the corresponding optical micrographs and relative positions between the cavity and the tapered fibre waveguide. Each of the eigenmodes is assigned a mode number (Fig. 4a) inferred from finite-element simulations (Methods). The importance of the phase matching between cavity and fibre eigenmodes is manifest in the observed transmission spectra. At larger gap widths ($d_g \approx 0.8, 0.4$ μm), only the resonances of the first- and second-order SPP eigenmodes (SPP_{1m} and SPP_{2m}) are observable, and the fundamental dielectric eigenmode (DE_{1m}) resonances are absent. This is because, for this range of larger gap widths, SPP eigenmodes are better phase-matched to the fibre eigenmode[26] and have a larger field overlap with the fundamental fibre eigenmode (they are located closer to the edge of, and extend farther outside, the microcavity than does the fundamental dielectric eigenmode in the wedge-shaped structure). As the gap width decreases further ($d_g \leq 0$), the fundamental dielectric eigenmodes are excited, as the phase-matching condition can be partly satisfied by decreasing d_g. For negative gap width, the SPP resonances are even more pronounced, as the phase-matching condition between the SPP and fibre eigenmodes can be fully satisfied owing to gap-width-induced phase matching, as is shown qualitatively in Fig. 2b. For the SPP resonance at 1565.4 nm, an input power transfer of up to 49.7% is demonstrated, showing the effectiveness of phase-matching control using the tapered fibre waveguide.

The demonstration of high-Q surface-plasmonic microcavities opens many possibilities for applications in fields ranging from fundamental science to device engineering. As a specific example, it could make possible a plasmonic laser, for which adequate gain materials as well as a high-Q SPP cavity are key prerequisites[27]. Although the demonstrated SPP Q factor is still less than that of an optical micro- or nanocavity[10,11], the corresponding SPP loss coefficient of $\alpha_{SPP} \approx 2\pi n_c/\lambda Q_{SPP} \approx 39\ cm^{-1}$ (where λ is the wavelength) satisfies the experimental criteria for a laser cavity and shows that, in principle, such surface-plasmonic lasing devices are possible. The tapered-fibre excitation scheme also demonstrates a convenient means of exciting these structures and selectively probing SPP cavity modes, because it directly controls the mode overlap and phase matching between the cavity and fibre eigenmodes (we also note

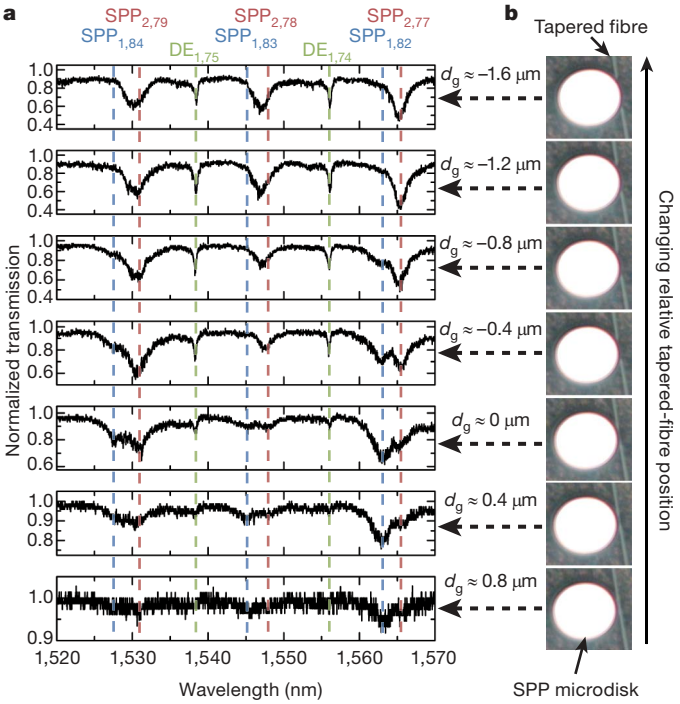


Figure 4 | Transmission spectrum versus waveguide coupling gap. a, Series of normalized transmission spectra, recorded for a variety of gap widths between the tapered fibre waveguide and the edge of the SPP microdisk. Resonances of SPP and dielectric eigenmodes are shown with estimated mode numbers. $R_b = 15.70 \pm 0.185\ \mu m$, $R_t = 13.08 \pm 0.14\ \mu m$, $d = 2\ \mu m$, $t \approx 100$ nm. For the SPP resonance at 1565.4 nm, an input power transfer of up to 49.7% is demonstrated (second panel from the top). **b**, Optical micrographs corresponding to the recorded normalized transmission spectra. Estimated gap width, d_g, is also shown.

that coupling to a conventional, chip-based waveguide is possible[28]). Furthermore, it is notable that the SPP Q factor could be substantially increased beyond the values measured here by lowering the temperature of the SPP microcavity[27,29]. From a fundamental standpoint, the SPP Q factor is sufficient to observe interesting cavity quantum electrodynamical phenomena in the weak-coupling regime relating to enhanced Purcell factors[11,29,30]. In addition, using the high nonlinearity of metal (or materials deposited in the vicinity of the metal), it may be possible to extend the applications of nonlinear plasmonics. Finally, it should be noted that, because the $\lambda^3 Q/V$ values of the present SPP microcavity (approximately a few hundred) are still much less than those provided by the photonic-crystal and dielectric whispering-gallery microcavities[10,11], it is still important to pursue new plasmonic cavity designs.

METHODS SUMMARY

Fabrication of SPP microdisk resonators. The template silica microdisks are fabricated by photolithography and buffered oxide etching as described elsewhere[19,21]. During the wet etch, the photoresist is undercut and produces a bevelled silica edge, which provides conformal silver coating of the top surface of the microdisk. The silver coating is deposited on the template silica microdisks using a d.c. sputtering technique with a chamber argon pressure of 30 mtorr. Two batches of samples (series 1 and 2) are prepared in this way to investigate the size-dependent characteristics of SPP microcavities.

Phase-matching condition. An asymptotic phase-matching formula can be obtained by two different approaches. Using the coupled-mode theory, the evaluation of the coupling coefficient κ involves the overlap integral of the cavity eigenmode and the tapered-fibre eigenmode. To have a non-zero coupling strength, the waveguide mode index can be approximated as written in equation (1) by setting the ϕ dependence of the integrand to zero. The same phase-matching condition can be found by path-averaging the effective mode index seen by the straight fibre waveguide[26]. This gives exactly the same formula

$$n_w \approx n_c \frac{2\tan^{-1}\left(\delta/\sqrt{\delta(2-\delta)}\right)}{\sqrt{\delta(2-\delta)}} = n_c\left(1+\frac{1}{3}\delta+\frac{2}{15}\delta^2+O(\delta^3)\right)$$

confirming the asymptotic dependence of phase matching on the relative gap width, δ. This formula applies only to the case of negative gap width, that is, $\delta = -d_g/R_b \geq 0$.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 12 July; accepted 6 November 2008.

1. Raether, H. R. *Surface Plasmons on Smooth and Rough Surfaces and on Gratings* (Springer, 1988).
2. Barnes, W. L., Dereux, A. & Ebbesen, T. W. Surface plasmon subwavelength optics. *Nature* **424,** 824–830 (2003).
3. Maier, S. A. *et al.* Local detection of electromagnetic energy transport below the diffraction limit in metal nanoparticle plasmon waveguides. *Nature Mater.* **2,** 229–232 (2003).
4. Stockman, M. I. Nanofocusing of optical energy in tapered plasmonic waveguides. *Phys. Rev. Lett.* **93,** 137404 (2004).
5. Ozbay, E. Plasmonics: Merging photonics and electronics at nanoscale dimensions. *Science* **311,** 189–193 (2006).
6. Cubukcu, E., Kort, E. A., Crozier, K. B. & Capasso, F. Plasmonic laser antenna. *Appl. Phys. Lett.* **89,** 093120 (2006).
7. Lopez-Tejeira, F. *et al.* Efficient unidirectional nanoslit couplers for surface plasmons. *Nature Phys.* **3,** 324–328 (2007).
8. Lal, S., Link, S. & Halas, N. J. Nano-optics from sensing to waveguiding. *Nature Photon.* **1,** 641–648 (2007).
9. Brongersma, M. L. & Kik, P. G. *Surface Plasmon Nanophotonics* (Springer, 2007).
10. Vahala, K. J. Optical microcavities. *Nature* **424,** 839–846 (2003).
11. Noda, S., Fujita, M. & Asano, T. Spontaneous-emission control by photonic crystals and nanocavities. *Nature Photon.* **1,** 449–458 (2007).
12. Ditlbacher, H. *et al.* Silver nanowires as surface plasmon resonators. *Phys. Rev. Lett.* **95,** 257403 (2005).
13. Bozhevolnyi, S. I., Volkov, V. S., Devaux, E., Laluet, J.-Y. & Ebbesen, T. W. Channel plasmon subwavelength waveguide components including interferometers and ring resonators. *Nature* **440,** 508–511 (2006).
14. Miyazaki, H. T. & Kurokawa, Y. Squeezing visible light waves into a 3-nm-thick and 55-nm-long plasmon cavity. *Phys. Rev. Lett.* **96,** 097401 (2006).
15. Weeber, J.-C., Bouhelier, A., Colas des Francs, G., Markey, L. & Dereux, A. Submicrometer in-plane integrated surface plasmon cavities. *Nano Lett.* **7,** 1352–1359 (2007).
16. Vesseur, E. J. R. *et al.* Surface plasmon polariton modes in a single-crystal Au nanoresonator fabricated using focused-ion-beam milling. *Appl. Phys. Lett.* **92,** 083110 (2008).
17. Cai, M., Painter, O. & Vahala, K. J. Observation of critical coupling in a fiber taper to a silica-microsphere whispering-gallery mode system. *Phys. Rev. Lett.* **85,** 74–77 (2000).
18. Spillane, S. M., Kippenberg, T. J., Painter, O. J. & Vahala, K. J. Ideality in a fiber-taper-coupled microresonator system for application to cavity quantum electrodynamics. *Phys. Rev. Lett.* **91,** 043902 (2003).
19. Kippenberg, T. J., Kalkman, J., Polman, A. & Vahala, K. J. Demonstration of an erbium-doped microdisk laser on a silicon chip. *Phys. Rev. A* **74,** 051802(R) (2006).
20. Borselli, M., Johnson, T. J. & Painter, O. Beyond the Rayleigh scattering limit in high-*Q* silicon microdisks: theory and experiment. *Opt. Express* **13,** 1515–1530 (2005).
21. Armani, D. K., Kippenberg, T. J., Spillane, S. M. & Vahala, K. J. Ultra-high-*Q* toroid microcavity on a chip. *Nature* **421,** 925–929 (2003).
22. Spillane, S. *et al.* Ultra-high-*Q* toroidal microcavities for cavity quantum electrodynamics. *Phys. Rev. A* **71,** 013817 (2005).
23. Oxborrow, M. Traceable 2-D finite-element simulation of the whispering-gallery modes of axisymmetric electromagnetic resonators. *IEEE Trans. Microw. Theory Tech.* **55,** 1209–1218 (2007).
24. Johnson, P. B. & Christy, R. W. Optical constants of noble metals. *Phys. Rev. B* **6,** 4370–4379 (1972).
25. Palik, E. D. *Handbook of Optical Constants of Solids* (Academic, 1985).
26. Rowland, D. R. & Love, J. D. Evanescent wave coupling of whispering gallery modes of a dielectric cylinder. *IEE Proc. J. Optoelectron.* **140,** 177–188 (1993).
27. Hill, M. T. *et al.* Lasing in metallic-coated nanocavities. *Nature Photon.* **1,** 589–594 (2007).
28. Almeida, V. R., Barrios, C. A., Panepucci, R. R. & Lipson, M. All-optical control of light on a silicon chip. *Nature* **431,** 1081–1084 (2004).
29. Gong, Y. Y. & Vučković, J. Design of plasmon cavities for solid-state cavity quantum electrodynamics applications. *Appl. Phys. Lett.* **90,** 033113 (2007).
30. Akimov, A. V. *et al.* Generation of single optical plasmons in metallic nanowires coupled to quantum dots. *Nature* **450,** 402–406 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank R. F. Oulton and G. Bartal for discussions and S. Zhang for a critical reading of the manuscript. This work was supported by the US Air Force Office of Scientific Research MURI program (grant no. FA9550-04-1-0434) and by the NSF Nanoscale Science and Engineering Center under award no. DMI-0327077.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to X.Z. (xiang@berkeley.edu) or K.V. (vahala@caltech.edu).

METHODS

Mode number of a dielectric eigenmode. The mode numbers, $h = 1, 2, \ldots$, of the dielectric eigenmodes DE_{hm} are assigned in order from lowest- to highest-order dielectric eigenmode. Depending on the geometry and the mode number h, dielectric eigenmodes can possess certain degrees of plasmonic characteristics due to the presence of the metal–silica interface.

Theoretical *Q*-factor estimation. From the finite-element eigenfrequency analysis, the complex-valued eigenfrequency, $f = f_{re} + if_{im}$, can be calculated, and Q factors evaluated using the formula $Q = f_{re}/2f_{im}$. The ranges of theoretical Q factors are estimated by using the error bounds in the imaginary part of the permittivity of silver. The radiation-limited Q factor can be estimated and separated from the metal-loss-limited Q factor by removing the imaginary part of the permittivity of silver. For example, the radiation-limited Q factor for $m = 54$ (Fig. 2c) is 3.9×10^6, and for $m = 85$ the Q factor is 6.7×10^9, both of which are orders of magnitude larger than the total Q factors.

Eigenmode identification. To assign mode numbers to the experimentally obtained resonance spectra, such as those shown in Fig. 4a, the size of the cavity is measured with a scanning electron microscope and the measured geometrical dimension is used in the finite-element calculation. Owing to the high sensitivity of the resonance frequency with respect to the nanoscale geometrical variation and the permittivity of the component materials, only the approximate mode numbers can be inferred. There being distinct ranges of Q factors indirectly confirms the theoretical SPP and dielectric resonance locations. Then the transmission of each resonance is experimentally determined by varying the gap width and the input polarization to assign distinct resonant characteristics precisely to each of the eigenmodes.

LETTERS

Warming of the Antarctic ice-sheet surface since the 1957 International Geophysical Year

Eric J. Steig[1], David P. Schneider[2], Scott D. Rutherford[3], Michael E. Mann[4], Josefino C. Comiso[5] & Drew T. Shindell[6]

Assessments of Antarctic temperature change have emphasized the contrast between strong warming of the Antarctic Peninsula and slight cooling of the Antarctic continental interior in recent decades[1]. This pattern of temperature change has been attributed to the increased strength of the circumpolar westerlies, largely in response to changes in stratospheric ozone[2]. This picture, however, is substantially incomplete owing to the sparseness and short duration of the observations. Here we show that significant warming extends well beyond the Antarctic Peninsula to cover most of West Antarctica, an area of warming much larger than previously reported. West Antarctic warming exceeds 0.1 °C per decade over the past 50 years, and is strongest in winter and spring. Although this is partly offset by autumn cooling in East Antarctica, the continent-wide average near-surface temperature trend is positive. Simulations using a general circulation model reproduce the essential features of the spatial pattern and the long-term trend, and we suggest that neither can be attributed directly to increases in the strength of the westerlies. Instead, regional changes in atmospheric circulation and associated changes in sea surface temperature and sea ice are required to explain the enhanced warming in West Antarctica.

Recent changes in Antarctic ice-sheet surface temperatures appear enigmatic when compared with global average temperature trends. Although the Antarctic Peninsula is one of the most rapidly warming locations on Earth, weather stations on the Antarctic continent generally show insignificant trends in recent decades[1]. However, all but two of the continuous records from weather stations are near the coast, providing little direct information on conditions in the continental interior. The widely used weather forecast reanalysis data are known to have errors owing to inconsistent assimilation skill in the satellite and pre-satellite eras[3].

In this Letter, we use statistical climate-field-reconstruction techniques to obtain a 50-year-long, spatially complete estimate of monthly Antarctic temperature anomalies. In essence, we use the spatial covariance structure of the surface temperature field to guide interpolation of the sparse but reliable 50-year-long records of 2-m temperature from occupied weather stations. Although it has been suggested that such interpolation is unreliable owing to the distances involved[1], large spatial scales are not inherently problematic if there is high spatial coherence, as is the case in continental Antarctica[4].

Previous reconstructions of Antarctic near-surface temperatures have yielded inconsistent results, particularly over West Antarctica, where records are few and discontinuous[5–7]. We improve upon this earlier work in several ways. We use two independent estimates of the spatial covariance of temperature across the Antarctic ice sheet: surface temperature measurements from satellite thermal infrared (T_{IR})

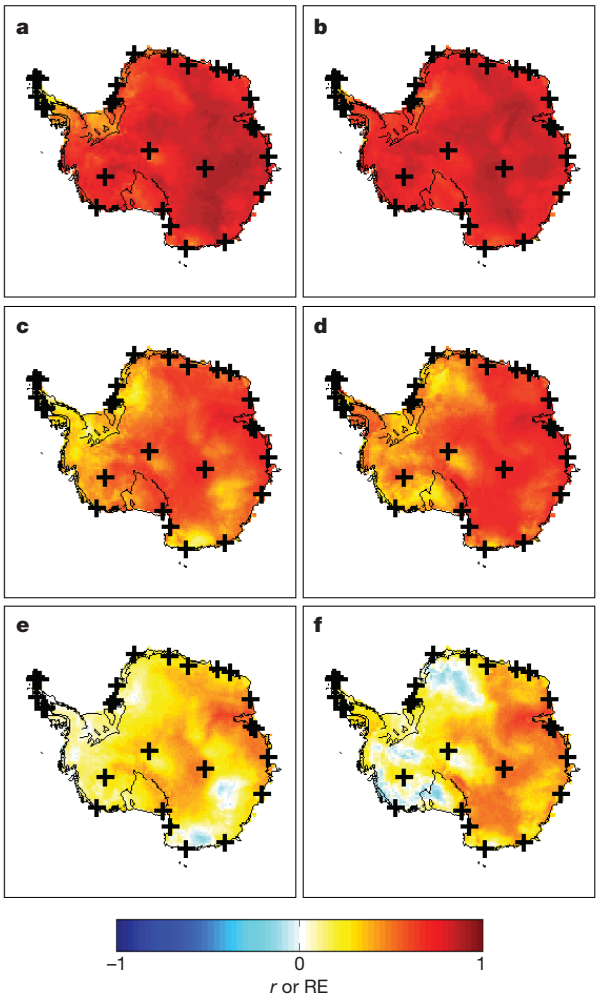


Figure 1 | Verification and upper-limit calibration statistics calculated for each grid point from the comparison of reconstructed and original satellite-derived monthly temperature anomalies. **a**, Calibration *r*, 1982–1994.5; **b**, calibration *r*, 1994.5–2006; **c**, verification *r*, 1994.5–2006; **d**, verification *r*, 1982–1994.5; **e**, verification RE, 1994.5–2006; **f**, verification RE, 1982–1994.5. Warm colours in **e** and **f** (RE scores greater than zero) show where results are more accurate than the climatological mean temperature. Mean grid-point verification results are RE = 0.11, CE = 0.09 and *r* = 0.46. Crosses show locations of occupied weather stations.

[1]Department of Earth and Space Sciences and Quaternary Research Center, University of Washington, Seattle, Washington 98195, USA. [2]National Center for Atmospheric Research, Boulder, Colorado 80307, USA. [3]Department of Environmental Science, Roger Williams University, Bristol, Rhode Island, USA. [4]Department of Meteorology, and Earth and Environmental Systems Institute, Pennsylvania State University, University Park, Pennsylvania 16802, USA. [5]NASA Laboratory for Hydrospheric and Biospheric Sciences, NASA Goddard Space Flight Center, Greenbelt, Maryland 20771, USA. [6]NASA Goddard Institute for Space Studies and Center for Climate Systems Research, Columbia University, New York, New York 10025, USA.

observations[8], and up-to-date automatic weather station (AWS) measurements of near-surface air temperature. We use a method[9,10] adapted from the regularized expectation maximization algorithm[11] (RegEM) for estimating missing data points in climate fields. RegEM is an iterative algorithm similar to principal-component analysis, used as a data-adaptive optimization of statistical weights for the weather station data. Unlike simple distance-weighting[5,6] or similar[7] calculations, application of RegEM takes into account temporal changes in the spatial covariance pattern, which depend on the relative importance of differing influences on Antarctic temperature at a given time. Furthermore, the iterative nature of RegEM allows it to be used with discontinuous time series, permitting us to take full advantage of the data available from occupied weather stations. We assess reconstruction skill using reduction-of-error (RE) and coefficient-of-efficiency (CE) scores as well as conventional correlation (r) scores. Such verification metrics are lacking in previous Antarctic temperature reconstructions[5–7], but are required for demonstrating skill relative to the climatological mean and are therefore critical for confidence in the calculation of temporal trends[10]. Skill metrics for our T_{IR}-based reconstruction from split calibration and verification experiments are significant (>99% confidence) at all grid points except in some restricted areas, mostly on the eastern side of the Antarctic Peninsula (Fig. 1).

Results from our AWS-based reconstruction agree well with those from the T_{IR} data (Fig. 2). This is important because the infrared data are strictly a measure of clear-sky temperature[8] and because surface temperature differs from air temperature 2–3 m above the surface, as measured at occupied stations or at AWSs. Trends in cloudiness or in the strength of the near-surface inversion could both produce spurious trends in the temperature reconstruction. The agreement between the reconstructions, however, rules out either potential bias as significant. Furthermore, detrending of the T_{IR} data before reconstruction demonstrates that the results do not depend strongly on trends in said data (Supplementary Information).

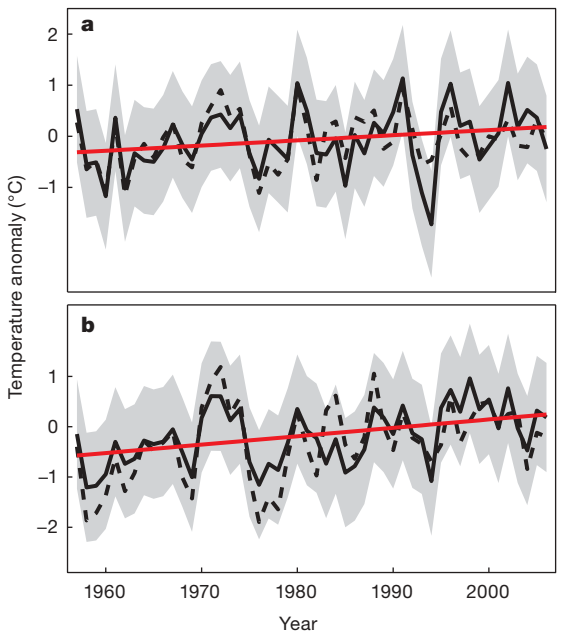


Figure 2 | Reconstructed annual mean Antarctic temperature anomalies, January 1957 to December 2006. **a**, East Antarctica; **b**, West Antarctica. Solid black lines show results from reconstruction using infrared satellite data, averaged over all grid points for each region. Dashed lines show the average of reconstructed AWS data in each region. Straight red lines show average trends of the T_{IR} reconstruction. Verification results for the continental mean of the T_{IR} reconstruction are RE = 0.34, CE = 0.31 and r = 0.73. Grey shading, 95% confidence limits.

Our reconstructions show more significant temperature change in Antarctica (Fig. 2), and a different pattern for that change than reported in some previous reconstructions[5,7] (Fig. 3). We find that West Antarctica warmed between 1957 and 2006 at a rate of 0.17 ± 0.06 °C per decade (95% confidence interval). Thus, the area of warming is much larger than the region of the Antarctic Peninsula. The peninsula warming averages 0.11 ± 0.04 °C per decade. We also find significant warming in East Antarctica at 0.10 ± 0.07 °C per decade (1957–2006). The continent-wide trend is 0.12 ± 0.07 °C per decade. In the reconstruction based on detrended T_{IR} data, warming in West Antarctica remains significant at greater than 99% confidence, and the continent-wide mean trend remains at 0.08 °C per decade, although it is no longer demonstrably different from zero (95% confidence). This is in good agreement with ref. 6, which reported average continent-wide warming of 0.082 °C per decade (1962–2003) and shows overall warming in West Antarctica, although statistical significance could not be demonstrated owing to the shorter length and greater variance of the reconstruction. We emphasize that, in general,

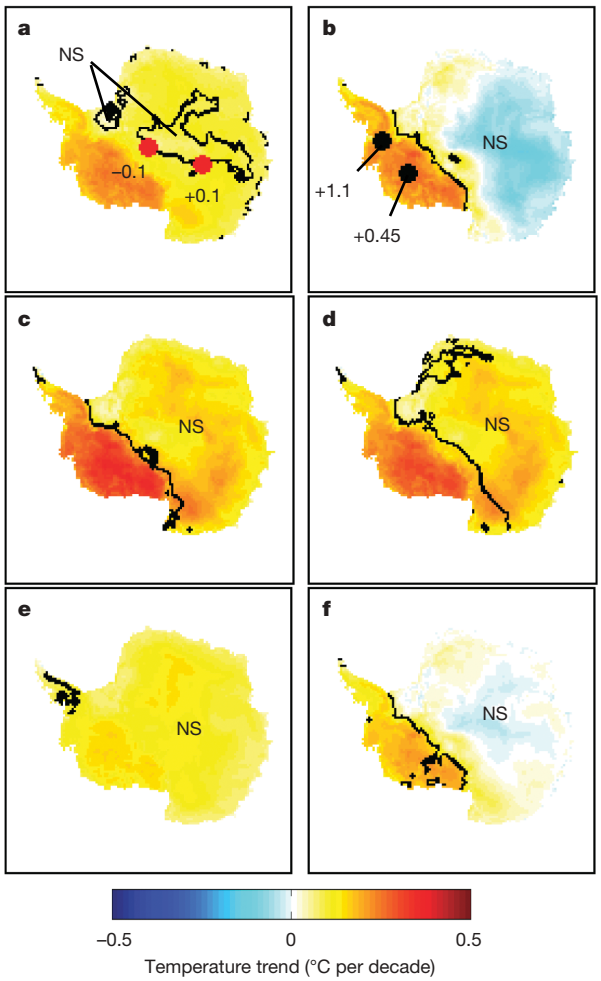


Figure 3 | Spatial pattern of temperature trends (degrees Celsius per decade) from reconstruction using infrared (T_{IR}) satellite data. **a**, Mean annual trends for 1957–2006; **b**, Mean annual trends for 1969–2000, to facilitate comparison with ref. 2. **c**–**f**, Seasonal trends for 1957–2006: winter (June, July, August; **c**); spring (September, October, November; **d**); summer (December, January, February; **e**); autumn (March, April, May; **f**). Black lines enclose those areas that have statistically significant trends at 95% confidence (two-tailed t-test). Where it would otherwise be unclear, NS (not significant) refers to areas of insignificant trends. Red circles and adjacent numbers in **a** show the locations of the South Pole and Vostok weather stations and their respective trends (degrees Celsius per decade) during the same time interval as the reconstruction (1957–2006). Black circles in **b** show the locations of Siple and Byrd Stations, and the adjacent numbers show their respective trends[13] for 1979–1997.

detrending of predictand data lowers the quality of reconstructions by removing spatial covariance information[10]. The detrended reconstruction therefore represents a conservative lower bound on trend magnitude. Although ref. 7 concluded that recent temperature trends in West Antarctica are statistically insignificant, the results were strongly influenced by the paucity of data from that region. When the complete set of West Antarctic AWS data is included, the trends become positive and statistically significant, in excellent agreement with our results[12].

Independent data provide additional evidence that warming has been significant in West Antarctica. At Siple Station (76° S, 84° W) and Byrd Station (80° S, 120° W), short intervals of data from AWSs were spliced with 37-GHz (microwave) satellite observations, which are not affected by clouds, to obtain continuous records from 1979 to 1997 (ref. 13). The results show mean trends of 1.1 ± 0.8 °C per decade and 0.45 ± 1.3 °C per decade at Siple and Byrd, respectively[13]. Our reconstruction yields 0.29 ± 0.26 °C per decade and 0.36 ± 0.37 °C per decade over the same interval. In our full 50-year reconstruction, the trends are significant, although smaller, at both Byrd (0.23 ± 0.09 °C per decade) and Siple (0.18 ± 0.06 °C per decade). Furthermore, the seasonal characteristics of these data[13] agree well with those from our reconstructions, with the greatest amount of warming in austral spring and winter (Fig. 3). Independent analyses of tropospheric temperature trends have also found spring and winter warming to be greatest in West Antarctica[14,15].

The spatial and seasonal characteristics of our temperature reconstruction have important implications for understanding recent Antarctic climate change. Several studies have emphasized a warming-peninsula, cooling-continent pattern that is attributed to changes in atmospheric circulation associated with the southern annular mode (SAM)[2,16]. Cooling over much of East Antarctica did occur in recent decades, but was strongest during the short time interval considered in earlier studies (1969–2000; Fig. 3b). Virtually all areas warmed between 1957 and ~1980. Our reconstruction differs from the results of modelling experiments that tie Antarctic surface temperature change to stratospheric ozone loss through changes in the SAM[16–18]. In such simulations, the largest negative temperature anomalies in East Antarctica occur in summer, whereas in our reconstruction, East Antarctic cooling is restricted to autumn (Fig. 3). The simulations show warming in austral summer and autumn, restricted to the peninsula, whereas in our reconstruction the greatest warming is in winter and spring, and in continental West Antarctica as well as on the peninsula.

The well-known increases in temperature on the Antarctic Peninsula are strongly associated with changes in sea ice[19]. Similarly, negative anomalies in sea-ice extent[20] and the length of the sea-ice season[21] in the Amundsen–Bellingshausen Sea may be related to the warming trends we observe in adjacent West Antarctica. To explore this, we examined model output from the NASA Goddard Institute for Space Studies (GISS) ModelE atmosphere-only and coupled general circulation models, which were run with multiple oceanic and atmospheric boundary conditions until the end of 2003 (ref. 22). A slightly earlier atmospheric version of GISS ModelE has been used in simulations of circulation anomalies associated with polar stratospheric ozone depletion[17]. When driven by observed sea-surface-temperature (SST) and sea-ice boundary conditions[23], the model reproduces many of the basic features of our reconstruction, with warming over most of the continent and persistent in West Antarctica (Fig. 4). SST and sea-ice changes alone produced weak cooling over parts of East Antarctica during the 1980s and 1990s. The details of the comparisons obviously depend on the accuracy of the SST and sea-ice observations (the latter are not generally considered reliable before 1979), and multi-decadal internal variability in the model is substantial. However, it is noteworthy that both in the reconstruction and in the model results, the rate of warming is greater in continental West Antarctica, particularly in spring and winter, than either on the peninsula or in East Antarctica. In GISS ModelE, this is related to SST changes and the location of sea-ice anomalies, particularly during the latter period (1979–2003), when they are strongly zonally asymmetric, with significant losses in the West Antarctic sector but small gains around the rest of the continent (Fig. 4e). Radiative forcings alone are inadequate to account for the observations (Supplementary Information).

The net impact of SST, sea ice, and radiative forcings on Antarctic temperatures in GISS ModelE is in general agreement with our reconstruction. The same model, when run in coupled mode (that is, with a dynamic ocean) fails to reproduce the strong trends observed in West Antarctica and the peninsula. The probable cause of this discrepancy, common to other coupled models[24], is inadequate representation of sea-ice anomalies and their associated higher-order modes of

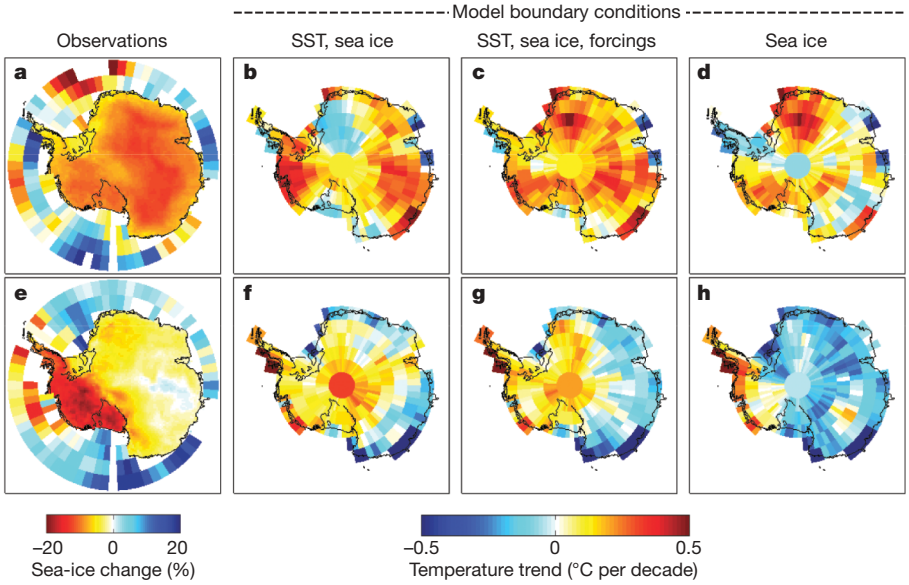


Figure 4 | Comparison of reconstructed and modelled mean annual temperature trends (degrees Celsius per decade) for the periods 1957–1981 and 1979–2003. **a–d**, 1957–1981; **e–h**, 1979–2003. **a**, **e**, Reconstructed surface temperature and observed 25-year change in sea-ice fractional area[23]. **b**, **f**, Surface air temperature from five-member GISS ModelE atmosphere-only ensemble simulations with observed sea-surface-temperature and sea-ice boundary conditions. **c**, **g**, Four-member ensemble with the same boundary conditions plus atmospheric forcings (changes in atmospheric concentrations of radiatively active species, including ozone). **d**, **h**, Difference between simulations with the same forcings but observed versus climatological sea ice, to isolate the effect of sea ice alone.

atmospheric circulation. In this context, it is important that the pattern of observed temperature trends closely resembles the pattern of temperature anomalies associated with the zonal wave-3 pattern in atmospheric circulation[4]. This circulation regime is efficient for the exchange of air between the ocean and the Antarctic continental interior, and is associated with atmospheric circulation anomalies in the Amundsen–Bellingshausen Sea, known to precede winter sea-ice anomalies[25]. Forced coupled models, including GISS ModelE, generally show a positive shift in the SAM and an associated increase in the circumpolar westerlies over recent decades, in good agreement with observations[26]. Observations also suggest a bias towards the positive phase in the wave-3 pattern[25] since about 1979, which is not reproduced in the coupled models. Using observed SST and sea ice, GISS ModelE does produce substantial shifts in the wave-3 circulation. Under those model conditions, greater cyclonic flow in the Amundsen Sea region brings warm, moist air to West Antarctica, countering the effect of the enhanced circumpolar westerlies.

An outstanding question in Antarctic climatology has been whether the strong warming of the peninsula has also occurred in continental West Antarctica[19]. Our results indicate that this is indeed the case, at least over the last 50 years. Moreover, ice-core analyses indicate average warming of West Antarctica over the entire twentieth century[27]. Although the influence of ozone-related changes in the SAM has been emphasized in recent studies of Antarctic temperature trends, the spatial and seasonal patterns of the observed temperature trends indicate that higher-order modes of atmospheric circulation, associated with regional sea-ice changes, have had a larger role in West Antarctica.

Mean surface temperature trends in both West and East Antarctica are positive for 1957–2006, and the mean continental warming is comparable to that for the Southern Hemisphere as a whole[28]. This warming trend is difficult to explain without the radiative forcing associated with increasing greenhouse-gas concentrations. However, the future trajectory of Antarctic temperature change also depends on the extent to which changes in atmospheric composition (whether from greenhouse gases or stratospheric ozone) affect Southern Hemisphere sea ice and regional atmospheric circulation patterns. Improved representation in models of coupled atmosphere/sea-ice dynamics will be critical for forecasting Antarctic temperature change.

METHODS SUMMARY

We use near-surface air temperature data from 42 occupied stations and 65 AWSs from the READER (Reference Antarctic Data for Environmental Research) data set[1]. We use passive infrared brightness measurements (T_{IR}) of surface temperature from the Advanced Very High Resolution Radiometer[8], a satellite of the US National Oceanic and Atmospheric Administration. We use the RegEM algorithm[9–11] to combine the data from occupied weather stations with the T_{IR} and AWS data in separate reconstructions of the near-surface Antarctic temperature field. Split calibration/verification tests are performed by withholding pre- and post-1995 T_{IR} and AWS data in separate RegEM calculations. Calibration and verification statistics are calculated for each grid point from the comparison of the reconstructed time series and the original temperature time series. We show RE and correlation r values in Fig. 1. CE verification values yield results indistinguishable from RE in our study and are reported in Supplementary Information. Significance levels of the calibration/verification statistics are based on Monte Carlo simulations of red noise as the null hypothesis. In Fig. 2, the 95% confidence interval is the unexplained variance, 2σ, where $\sigma^2_{error} = \sigma^2_{data}(1 - r^2_{ver})$, σ^2_{data} is the temporal variance in the original satellite temperature data and r^2_{ver} is the verification fractional resolved variance. Significance levels of trends are calculated using a two-tailed t-test, with the number of degrees of freedom adjusted for autocorrelation. In reporting trends for different areas, we define West Antarctica as 72°–90° S, 60°–180° W; East Antarctica as 65°–90° S, 300–180° E; and the Antarctic Peninsula as westerly longitudes north of 72° S.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 14 January; accepted 1 December 2008.

1. Turner, J. *et al.* Antarctic climate change during the last 50 years. *Int. J. Climatol.* **25,** 279–294 (2005).
2. Thompson, D. W. J. & Solomon, S. Interpretation of recent Southern Hemisphere climate change. *Science* **296,** 895–899 (2002).
3. Bromwich, D. H. & Fogt, R. L. Strong trends in the skill of the ERA-40 and NCEP-NCAR Reanalyses in the high and midlatitudes of the southern hemisphere, 1958–2001. *J. Clim.* **17,** 4603–4619 (2004).
4. Schneider, D. P., Steig, E. J. & Comiso, J. Recent climate variability in Antarctica from satellite-derived temperature data. *J. Clim.* **17,** 1569–1583 (2004).
5. Doran, P. T. *et al.* Antarctic climate cooling and terrestrial ecosystem response. *Nature* **415,** 517–520 (2002).
6. Chapman, W. L. & Walsh, J. E. A synthesis of Antarctic temperatures. *J. Clim.* **20,** 4096–4117 (2007).
7. Monaghan, A. J., Bromwich, D. H., Chapman, W. & Comiso, J. C. Recent variability and trends of Antarctic near-surface temperature. *J. Geophys. Res.* **113,** doi:1029/2007JD009094 (2008).
8. Comiso, J. C. Variability and trends in Antarctic surface temperatures from in situ and satellite infrared measurements. *J. Clim.* **13,** 1674–1696 (2000).
9. Rutherford, S. *et al.* Proxy-based Northern Hemisphere surface temperature reconstructions: Sensitivity to methodology, predictor network, target season and target domain. *J. Clim.* **18,** 2308–2329 (2005).
10. Mann, M. E., Rutherford, S., Wahl, E. & Ammann, C. Robustness of proxy-based climate field reconstruction methods. *J. Geophys. Res.* **112,** doi:10.1029/2006JD008272 (2007).
11. Schneider, T. Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *J. Clim.* **14,** 853–871 (2001).
12. Bromwich, D. H., Monaghan, A. J. & Colwell, S. R. Surface and Mid-tropospheric Climate Change in Antarctica. *Eos* **89** (Fall meeting), abstr. C41A-0497 (2008).
13. Shuman, C. A. & Stearns, C. R. Decadal-length composite inland West Antarctic temperature records. *J. Clim.* **14,** 1977–1988 (2001).
14. Johanson, C. M. & Fu, Q. Antarctic atmospheric temperature trend patterns from satellite observations. *Geophys. Res. Lett.* **34,** doi:1029/2006GL029108 (2007).
15. Turner, J. *et al.* Significant warming of the Antarctic winter troposphere. *Science* **311,** 1914–1917 (2006).
16. Gillett, N. P. & Thompson, D. W. J. Simulation of recent Southern Hemisphere climate change. *Science* **302,** 273–275 (2003).
17. Shindell, D. T. & Schmidt, G. A. Southern Hemisphere climate response to ozone changes and greenhouse gas increases. *Geophys. Res. Lett.* **31,** doi:10.1029/2004GL020724 (2004).
18. Keeley, S. P. E. *et al.* Is Antarctic climate most sensitive to ozone depletion in the middle or lower stratosphere? *Geophys. Res. Lett.* **34,** doi:10.1029/2007GL031238 (2007).
19. Vaughan, D. G. *et al.* Recent rapid regional climate warming on the Antarctic Peninsula. *Clim. Change* **60,** 243–274 (2003).
20. Kwok, R. & Comiso, J. C. Southern Ocean climate and sea ice anomalies associated with the Southern Oscillation. *J. Clim.* **15,** 487–501 (2002).
21. Parkinson, C. L. Trends in the length of the Southern Ocean sea ice season, 1979–1999. *Ann. Glaciol.* **34,** 435–440 (2002).
22. Hansen, J. *et al.* Climate simulations for 1880–2003 with GISS Model E. *Clim. Dyn.* **29,** 661–696 (2007).
23. Rayner, N. A. *et al.* Global analyses of sea surface temperature, sea ice and night marine air temperatures since the late nineteenth century. *J. Geophys. Res.* **108,** doi:10.1029/2002JD002670 (2003).
24. Connolley, W. M. & Bracegirdle, T. J. An Antarctic assessment of IPCC AR4 coupled models. *Geophys. Res. Lett.* **34,** doi:10.1029/2007GL031648 (2007).
25. Holland, M. M. & Raphael, M. Twentieth century simulation of the Southern Hemisphere in coupled models. Part II: Sea ice conditions and variability. *Clim. Dyn.* **26,** 229–245 (2006).
26. Miller, R. L., Schmidt, G. A. & Shindell, D. T. Forced annular variations in the 20th Century Intergovernmental Panel on Climate Change Fourth Assessment Report models. *J. Geophys. Res.* **111,** doi:10.1029/2005JD006323 (2006).
27. Schneider, D. P. & Steig, E. J. Ice cores record significant 1940s Antarctic warmth related to tropical climate variability. *Proc. Natl Acad. Sci. USA* **105,** 12154–12158 (2008).
28. Jones, P. D. & Moberg, A. Hemispheric and large-scale surface air temperature variations: An extensive revision and an update to 2001. *J. Clim.* **16,** 206–223 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements E.J.S. and D.P.S. were supported by the US National Science Foundation, grant numbers OPP-0440414 and OPP-0126161, as part of the US ITASE programme. M.E.M. was supported by the US National Science Foundation, grant number OPP-0125670. We thank D. Winebrenner, A. Monaghan, D. Bromwich, J. Turner, P. Mayewski, T. Scambos, E. Bard and O. Bellier.

Author Contributions E.J.S., D.P.S., S.D.R. and M.E.M. made the reconstruction and statistical calculations. J.C.C. performed the cloud-masking calculations and provided the updated satellite data set. D.T.S. provided the general circulation model output and guided its interpretation. E.J.S. wrote the paper. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to E.J.S. (steig@ess.washington.edu).

METHODS

Data. We use the READER weather station temperatures from the British Antarctic Survey[1]. Twenty-seven of the 42 occupied stations have at least 50%-complete monthly average data from 1957 to present. Data from 65 AWSs are available, but are discontinuous and date from 1980 at the earliest. In addition, data from only 24 of the AWSs are more than 50% complete for 1980–2006. We use passive infrared brightness measurements (T_{IR}) from the Advanced Very High Resolution Radiometer, which are continuous beginning January 1982 and constitute the most spatially complete Antarctic temperature data set. The T_{IR} data are biased towards clear-sky conditions, owing to the opacity of clouds in the infrared band. Cloud masking is probably the largest source of error in the retrieval of T_{IR} data from raw satellite spectral information. We have updated the data throughout 2006, using an enhanced cloud-masking technique to give better fidelity with existing occupied and automatic weather station data. We make use of the cloud masking in ref. 8 but impose an additional restriction that requires that daily anomalies be within a threshold of $\pm 10\,^{\circ}$C of climatology, a conservative technique that will tend to damp extreme values and, hence, minimize trends[29]. Values that fall outside the threshold are removed.

Calculations. We use the RegEM algorithm[11], developed for sparse data infilling, to combine the occupied weather station data with the T_{IR} and AWS data in separate reconstructions of the Antarctic temperature field. RegEM uses an iterative calculation that converges on reconstructed fields that are most consistent with the covariance information present both in the predictor data (in this case the weather stations) and the predictand data (the satellite observations or AWS data). We use an adaptation of RegEM in which only a small number, k, of significant eigenvectors are used[10]. Additionally, we use a truncated total-least-squares (TTLS) calculation[30] that minimizes both the vector $\mathbf{b}$ and the matrix A in the linear regression model $A\mathbf{x} = \mathbf{b}$. (In this case A is the space-time data matrix, $\mathbf{b}$ is the principal component time series to be reconstructed and $\mathbf{x}$ represents the statistical weights.) Using RegEM with TTLS provides more robust results for climate field reconstruction than the ridge-regression method originally suggested in ref. 11 for data infilling problems, when there are large differences in data availability between the calibration and reconstruction intervals[10]. For completeness, we compare results from RegEM with those from conventional principal-component analysis (Supplementary Information).

Monthly average surface temperature anomalies were obtained from the T_{IR} data for the domain covering all land areas and ice shelves on the Antarctic continent, at 50 km × 50 km resolution[4]. The monthly anomalies are efficiently characterized by a small number of spatial weighting patterns and corresponding time series (principal components) that describe the varying contribution of each pattern. The results are reproducible using single-season, annual average and split-decade-length subsets of the data[4]. The first three principal components are statistically separable and can be meaningfully related to important dynamical features of high-latitude Southern Hemisphere atmospheric circulation, as defined independently by extrapolar instrumental data. The first principal component is significantly correlated with the SAM index (the first principal component of sea-level-pressure or 500-hPa geopotential heights for 20° S–90° S), and the second principal component reflects the zonal wave-3 pattern, which contributes to the Antarctic dipole pattern of sea-ice anomalies in the Ross Sea and Weddell Sea sectors[4,8]. The first two principal components of T_{IR} alone explain >50% of the monthly and annual temperature variabilities[4]. Monthly anomalies from microwave data (not affected by clouds) yield virtually identical results[4].

Principal component analysis of the weather station data produces results similar to those of the satellite data analysis, yielding three separable principal components. We therefore used the RegEM algorithm with a cut-off parameter $k = 3$. A disadvantage of excluding higher-order terms ($k > 3$) is that this fails to fully capture the variance in the Antarctic Peninsula region. We accept this trade-off because the Peninsula is already the best-observed region of the Antarctic.

Statistics. We obtained calibration/verification statistics by withholding the first and last 12.5 years of the 25-year T_{IR} data in separate RegEM calculations. We similarly split the AWS data into pre- and post-1995 data. Confidence levels are based on Monte Carlo simulations of red noise as the null hypothesis. For each grid point, 1,000 red noise series were generated to have the same mean, variance and lag-1 autocorrelation coefficient as the actual time series over the calibration period. Additional validation of the T_{IR}-based reconstruction was obtained by using the 15 occupied weather stations with the most complete data, reserving the other 27 for verification. Verification metrics at these sites are consistently significant at >99% confidence, with the exception of some sites at the tip of the Antarctic Peninsula and the three sites north of 55° S (Supplementary Information).

We report verification statistics as well as upper-bound calibration-interval statistics. The latter represent the maximum level of explained variance that could be expected in the reconstruction, given how much data variance is resolved over the calibration interval. We rely primarily on the RE statistic; the alternative verification statistic, CE, yields indistinguishable results in our study (Supplementary Information). For completeness, we also report correlation r values, but with the recognition that r is a deficient skill metric because it does not penalize the poor prediction of either means or variances[10].

29. Reynolds, R. W. *et al.* An improved *in situ* and satellite SST analysis for climate. *J. Clim.* **15,** 1609–1625 (2002).
30. Fierro, R. D., Golub, G. H., Hansen, P. C. & O'Leary, D. P. Regularization by truncated total least squares. *SIAM J. Sci. Comput.* **18,** 1223–1241 (1997).

LETTERS

A simple model of bipartite cooperation for ecological and organizational networks

Serguei Saavedra[1,2,3], Felix Reed-Tsochas[2,4] & Brian Uzzi[5,6]

In theoretical ecology, simple stochastic models that satisfy two basic conditions about the distribution of niche values and feeding ranges have proved successful in reproducing the overall structural properties of real food webs, using species richness and connectance as the only input parameters[1–4]. Recently, more detailed models have incorporated higher levels of constraint in order to reproduce the actual links observed in real food webs[5,6]. Here, building on previous stochastic models of consumer–resource interactions between species[1–3], we propose a highly parsimonious model that can reproduce the overall bipartite structure of cooperative partner–partner interactions, as exemplified by plant–animal mutualistic networks[7]. Our stochastic model of bipartite cooperation uses simple specialization and interaction rules, and only requires three empirical input parameters. We test the bipartite cooperation model on ten large pollination data sets that have been compiled in the literature, and find that it successfully replicates the degree distribution, nestedness and modularity of the empirical networks. These properties are regarded as key to understanding cooperation in mutualistic networks[8–10]. We also apply our model to an extensive data set of two classes of company engaged in joint production in the garment industry. Using the same metrics, we find that the network of manufacturer–contractor interactions exhibits similar structural patterns to plant–animal pollination networks. This surprising correspondence between ecological and organizational networks suggests that the simple rules of cooperation that generate bipartite networks may be generic, and could prove relevant in many different domains, ranging from biological systems to human society[11–14].

In ecology, the collection and analysis of empirical data on mutualistic networks[7] can help identify the most significant features that have evolved in bipartite cooperative networks. In these networks, species in one class (class *A*, for animals) cooperate with species in a second class (class *P*, for plants) to mutual advantage. Species in class *P* offer rewards with certain characteristics determined by reward traits, which may also have evolved to reduce exploitation and favour mutualism[15]. Species in class *A* foraging for resources can benefit from the rewards offered by a given species in class *P* if the respective foraging traits (for example efficiency, morphology and behaviour) and reward traits (for example quantity, quality and availability) are complementary[8,16]. External factors such as the environmental context (for example population density and geographic and temporal variation) attenuate or amplify the value of reward and foraging traits, and have an impact on the number of potential partners with which a given species cooperates[17–19]. Furthermore, mutualistic networks exhibit constraints introduced by phylogenetic relationships between species in the same class, which impact mutualistic interaction patterns by favouring ecological similarity[20].

Recently, mutualistic models[21,22] have been proposed which incorporate the effects of complementarity and exploitation or growth barriers on interactions between species in classes *A* and *P* (Supplementary Information). Although a mixed model[21] and a differential limiting size model[22] have proved successful at generating structural patterns similar to those observed in real mutualistic networks, we find that they are able to reproduce less than 25% of the total number of observed metrics in different ecological networks (Supplementary Tables 1–4).

In organizational networks, cooperation between two different classes of company (manufacturers and contractors) is subject to equivalent structural constraints, which depend on the traits of the companies and the complementarity between traits of potential partners, as well as hierarchical relationships between companies in the same class[23]. Companies are characterized by a set of organizational traits[24] (for example company size, competitive niche space and brand positioning), and face interaction barriers generated by differences in status[25], which limit the number and range of potential partners. Again, the values associated with trait complementarity and interaction barriers are not absolute, but are modulated by the specific market context in which the companies operate[24,25].

Inspired by previous food-web models[1–3], we develop a new model of bipartite cooperation that can reproduce more than 70% of the total number of observed metrics in both ecological and organizational networks (Table 1, Supplementary Tables 3, 4). In the corresponding ecological (species-based) and organizational (company-based) contexts, plants and manufacturers are treated as members of class *P*, and animals and contractors as members of class *A*. The three inputs for the model are the size of class *A*, the size of class *P* and the total number of links, *L*, all of which are given directly by the empirical data (Table 1). The model consists of two mechanisms (see Methods for a description of the model).

The first mechanism is specialization. The specialization rule determines how many partners, l_p, each member $p \in P$ will cooperate with. This number is determined by the reward value associated with $p \in P$, which is given by the reward trait, t_{Rp}, attenuated or amplified by an external factor λ_p that accounts for effects such as geographic variation and population diversity. Higher reward values increase the number of potential interactions. Reward traits t_{Rp} are the result of a hierarchical process, which in our model corresponds to the generation of an ordered sequence in trait space, so t_{Rp} has a role equivalent to that of the niche value in the niche model[1] or nested-hierarchy model[2]. Note that the specialization rule is only associated with class *P*. This is consistent with previous findings which show that external factors affect the level of specialization more strongly in plants than in animals[18].

The second mechanism is interaction. The interaction rule determines which members $a \in A$ cooperate with each member $p \in P$. Here interactions are limited by the complementarity between the reward traits, t_{Rp}, for $p \in P$ and foraging traits (organizational traits), t_{Fa}, for $a \in A$. The foraging trait t_{Fa} limits the range of possible partners for each member of class *A*, but again external factors λ_{lp} such as temporal variation and population density can modify these

[1]Department of Engineering Science, University of Oxford, Oxford OX1 3PJ, UK. [2]CABDyN Complexity Centre, [3]Corporate Reputation Centre, [4]James Martin Institute, Saïd Business School, University of Oxford, Oxford OX1 1HP, UK. [5]Kellogg School of Management and Northwestern Institute on Complex Systems, Northwestern University, Evanston, Illinois 60208, USA. [6]Haas School of Business, University of California, Berkeley, California 94720, USA.

Table 1 | Empirical values and model statistical significance

Data set/environment	L	$\|P\|$	$\|A\|$	KS_P–KS_A	N	Q
Marsh (Japan)	430	64	187	0.326*–0.438*	0.976* (0.969)	0.551* (0.553)
Grassland (Cass, New Zealand)	374	41	139	0.633*–0.385*	0.957* (0.960)	0.474† (0.465)
Subalpine forest/meadow (Japan)	865	90	354	0.552*–0.001§	0.985† (0.976)	0.545† (0.532)
Subalpine (Arthur's Pass, New Zealand)	120	18	60	0.108†–0.999*	0.858§ (0.936)	0.553‡ (0.527)
Subalpine (Craigieburn, New Zealand)	346	49	118	0.002§–0.001§	0.961* (0.955)	0.480† (0.468)
Tundra (Canada)	179	29	81	0.097†–0.989*	0.971† (0.950)	NM
Scrub/snow gum forest (Australia)	252	36	81	0.608*–0.076†	0.935* (0.949)	NM
Deciduous forest (USA)	65	7	33	0.911*–0.642*	0.953† (0.930)	NM
Arctic tundra (Greenland)	453	31	75	0.038‡–0.118†	0.793§ (0.914)	NM
Subarctic rock slope (Sweden)	242	24	118	0.223†–0.005‡	0.927† (0.952)	NM
New York garment industry, 1985	7,250	823	2,562	0.061†–0.115†	0.997† (0.996)	0.598§ (0.502)
New York garment industry, 1991	3,981	325	1,590	0.101†–0.531i	0.994* (0.993)	0.601§ (0.529)
New York garment industry, 1997	1,450	148	700	0.003‡–0.264†	0.990† (0.988)	0.653‡ (0.625)
New York garment industry, 2003	228	62	128	0.370†–0.002‡	0.976† (0.969)	0.711† (0.700)

For each pollination data set and the four organizational networks used in this paper, the table presents its environment/location; the total number of links, L; the respective numbers of nodes, $|P|$ and $|A|$, in classes P and A; the combined Kolmogorov–Smirnov (KS) probabilities, KS_P–KS_A, calculated for the degree distributions using the two-group equivalence KS test between the empirical and model-generated distributions for classes P and A, respectively; the observed nestedness, N; and the observed mean modularity value, Q. Note that all networks have a ratio $|P|/|A| < 0.5$, which has been found to be an important factor limiting the appearance of scale-free distributions[22]. The model-generated mean values for N and Q are shown inside parentheses. Five of the observed pollination networks have already been found to be non-modular (NM)[10]. All comparisons are based on 1,000 model simulations. Note that the model reproduces more than 70% of the overall number of observed metrics with a good or excellent fit (27 out of 35 and 11 out of 16 for the ecological and organizational networks, respectively).
* KS ≥ 0.30, normalized errors less than one model s.d. (excellent fit). † KS < 0.30, normalized errors between one and two model s.d. (good fit). ‡ KS < 0.05, normalized errors between two and three model s.d. (poor fit). § KS < 0.01, normalized errors greater than three model s.d. (bad fit).

interaction barriers. In both the specialization and interaction rules, we assume that traits are normally distributed and that the effect of external factors on members is inhomogeneous and a function of the diversity of members and the density of interactions in the network.

We test our model using data from real-world ecological and organizational networks. For the ecological networks, we use a diverse set of ten extensive plant–animal pollination networks compiled in the literature (Supplementary Information), which can clearly be distinguished from random assemblages[9]. For organizational networks, we use a data set of approximately 700,000 yearly bilateral production transactions between more than 10,000 manufacturers and contractors in the New York garment industry (NYGI) from January 1985 to December 2003 (Supplementary Information)[23]. The NYGI exhibits a significant turnover of companies each year with different declining trajectories for each class (Supplementary Fig. 1), so the network structure at one time does not trivially map into the structure at other times. For each network, we investigate three key features of bipartite cooperation[8–10]: (1) degree distribution[26], (2) nestedness[27,28] and (3) modularity[29].

First we consider the degree distribution. In Fig. 1a and b we respectively show the cumulative distributions for members of class P

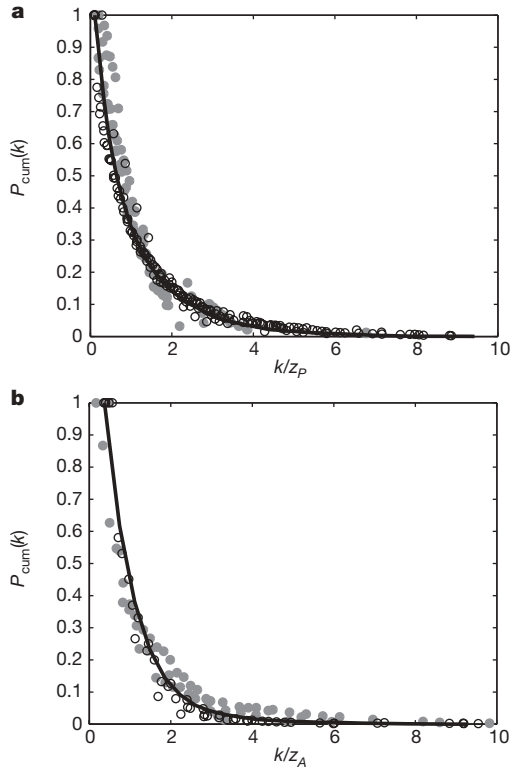


Figure 1 | Scaled degree distribution. **a**, The cumulative degree distribution, $P_{cum}(k)$, for members of class P (plants, manufacturers); **b**, $P_{cum}(k)$ for members of class A (animals, contractors). The number of partners, k, is scaled by a multiplicative factor of $1/z_P$ for members of P and $1/z_A$ for members of A, where $z_P = L/P$ and $z_A = L/A$. Filled symbols correspond to pollination networks and open symbols to organizational networks. Note that all distributions collapse into a single curve. The solid line corresponds to the model-generated distributions averaged over 1,000 simulations (Supplementary Fig. 2).

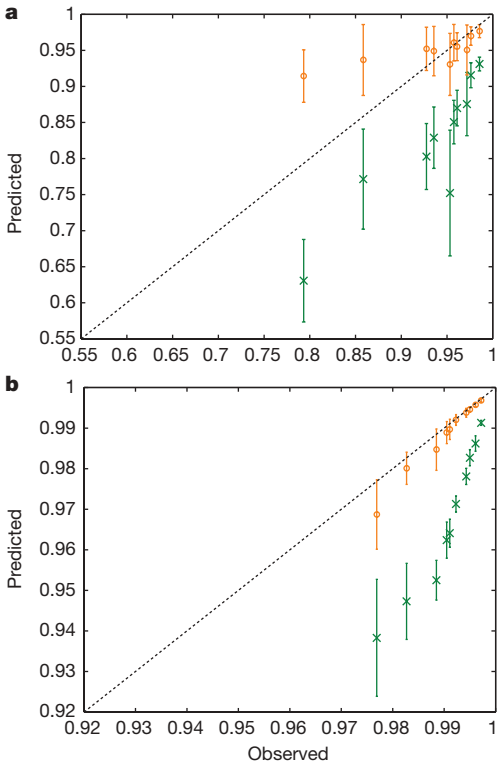


Figure 2 | Nestedness. **a**, **b**, Comparisons of the randomly generated (green crosses) and model-generated (orange circles) nestedness values (mean ± 2 s.d.) with the observed nestedness values for all the pollination (**a**) and the NYGI (**b**) networks. The dashed lines correspond to there being perfect agreement between predicted and observed values. Note that a matrix with a nestedness value of one is perfectly nested.

and members of class *A*. Note that pollination networks (filled symbols) and organizational networks (open symbols) exhibit the same patterns in both degree distributions. The solid line corresponds to the model-generated degree distributions (see also Supplementary Fig. 2). Table 1 provides statistical evidence for the correspondence between the empirical and model-generated distributions for both ecological and organizational networks.

To calculate the nestedness level, *N*, of the empirical and model-generated networks, we use the BINMATNEST program[28]. Here nestedness is defined in the interval [0, 1], where 1 corresponds to a perfectly nested network[9]. Figure 2 shows the nestedness values for empirical networks (dashed line), random assemblages generated following the null model II proposed in ref. 9 (green crosses) and model-generated networks (orange circles). Note that the model always significantly outperforms random assemblages. Table 1 presents tests of statistical significance and shows that there is a high level of correspondence between the model and data for most of the networks (Supplementary Fig. 3).

Modularity values, *Q*, for the networks are calculated using the one-mode optimization algorithm proposed in ref. 30, because we want to extract only cooperative units from the network. This algorithm has shown that only five of the observed ecological networks are truly modular[10] (Table 1). For these five networks, we find a strong correspondence between the empirical and model-generated modularity values (Table 1). For each organizational network, we find a modularity value that is higher than that generated from the corresponding random assemblages[30] ($P < 10^{-6}$), and this empirical behaviour is replicated by the model (Table 1). Nodes have been shown to have different connectivity roles (for example global or local hubs) depending on how they are embedded within their module and participate in other modules[30]. Here we also test the ability of our model to accurately reproduce the empirically observed number of nodes within each connectivity role[30] (Fig. 3). To do so, we measure the Pearson correlation coefficient, *r*, and the ratio of the connectivity role norms, *d*, for the observed and model-generated networks (Methods). We consistently find values aligned with the empirical measurements for the ecological and organizational networks ($r = 0.98$, $0.9 < d < 1.1$).

Our study identifies striking similarities in the general structural characteristics of networks that are formed as a result of cooperative mechanisms operating in radically different contexts, linking partners in ecological and socio-economic systems, respectively. This empirical finding motivates the proposed simple model for bipartite cooperation, which captures the most important generic features of mutualistic interaction patterns starting from a minimal set of input parameters. At the level of partner–partner interactions, equivalent behaviour in different systems appears to be driven by similar types of interaction constraints. These correspond to complementarity in traits or characteristics, a hierarchical organization limiting the range

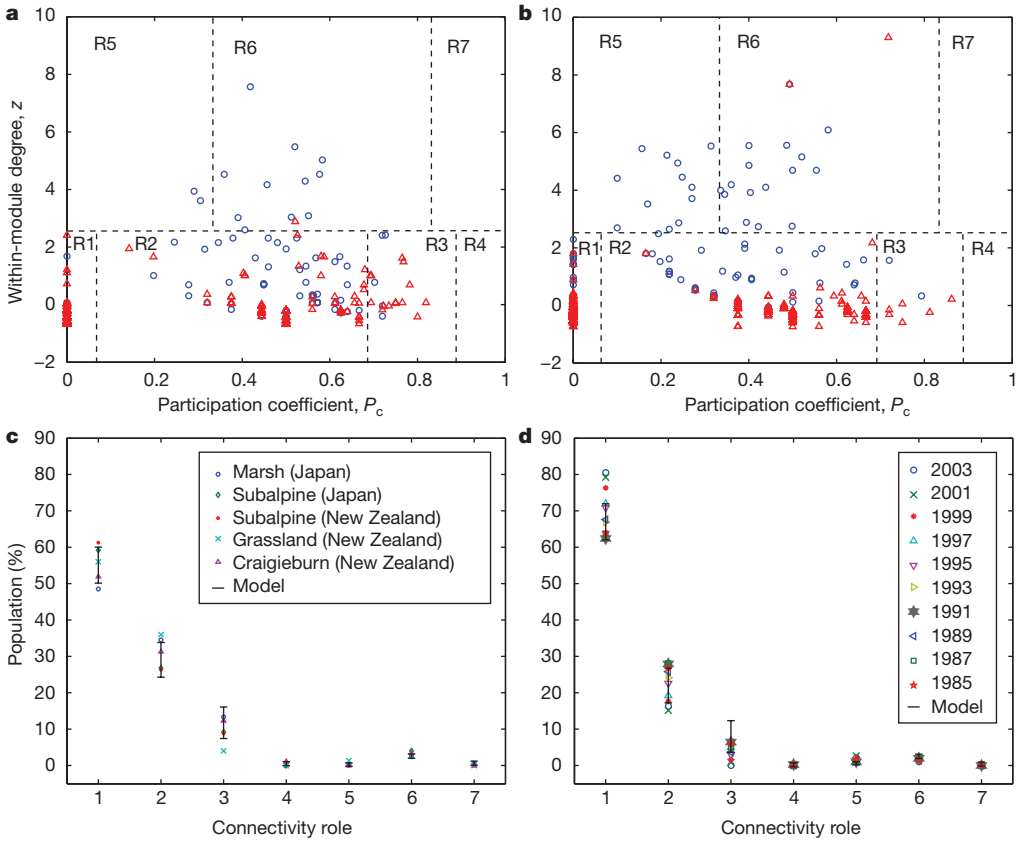


Figure 3 | Connectivity roles. **a**, **b**, Division of connectivity roles (R1–R7) for members of the subalpine (Japan) network (**a**) and the 1997 NYGI network (**b**). Plants and manufacturers are indicated with blue circles and animals and contractors with red triangles. Nodes with a normalized within-module degree (*z* score) $z \geq 2.5$ have been heuristically defined as module hubs, and nodes with $z < 2.5$ as non-hub nodes[30]. In addition, hubs and non-hub nodes are classified according to a participation coefficient, P_c, which gives the level of interaction of a node with the rest of the modules. The complete classification is as follows. For non-hub nodes, R1 comprises nodes with all their links connected within their own module ($P_c \leq 0.05$), R2 comprises nodes with most of their links connected within their own module ($0.05 < P_c \leq 0.62$), R3 comprises nodes with many links connected to other modules ($0.62 < P_c \leq 0.80$) and R4 comprises nodes with their links homogeneously connected to all other modules ($P_c > 0.80$). For hub-nodes, R5 comprises nodes with most of their links connected to their own module ($P_c \leq 0.30$), R6 comprises nodes with most of their links connected to other modules ($0.30 < P_c \leq 0.75$) and R7 comprises nodes with their links homogeneously connected to all other modules ($P_c > 0.75$). **c**, **d**, Proportions of nodes from the population within each connectivity role for the modular pollination networks (**c**) and the NYGI networks (**d**). Points (various symbols) correspond to the empirical values and bars (mean ± s.d.) correspond to the model-generated values. Note that the model accurately replicates the proportion of nodes within each connectivity role for both types of network.

of potential partners, and the environmental context. The success of this simple stochastic model in generating the overall structural characteristics of mutualistic networks makes it a suitable starting point for the development of more elaborate ecological models, with the aim of addressing the important question of reproducing the actual links observed in real mutualistic webs. Such an approach would require more comprehensive statistical comparisons across different network metrics using maximum-likelihood methods[6]. Beyond this ecological context, and the specific organizational network that has been considered, the generic nature of the bipartite cooperation model and its underlying assembly rules suggests that it should be relevant to many other cooperative networks in domains beyond those considered here.

METHODS SUMMARY

Model. For the specialization rule, the number of links, l_p, for each node $p \in P$ is defined by $lp = 1 + \mathrm{Round}((L-|P|)t_{\mathrm{R}p}\lambda_p / \sum_j t_{\mathrm{R}j}\lambda_j)$, where the reward trait $t_{\mathrm{R}p}$ is uniformly drawn from $[0, 1]$, the external factor λ_p is randomly drawn from an exponential distribution, $|\cdot|$ denotes set cardinality and $\mathrm{Round}(\cdot)$ is the nearest-integer function.

For the interaction rule, nodes from class P are sorted according to their reward trait $t_{\mathrm{R}p}$ in ascending order. Nodes from class A are sorted in descending order according to their foraging traits $t_{\mathrm{F}a}$, which are uniformly drawn from $[0, 1]$. Starting from the first node, p_i, and continuing sequentially subject to $t_{\mathrm{R}pi} > \lambda_{lpi}$, each link l_{pi} is connected to the first node $a' \in A'$, where A' is the subset of nodes in A that have not already been linked to by another node $p \neq p_i$. Here λ_{lpi} is an external factor drawn randomly from the same exponential distribution as λ_{pi}. If $t_{\mathrm{R}pi} \leq \lambda_{lpi}$ then the link is randomly connected to another node $a'' \in A''$, where A'' is the subset of nodes in A that have been linked in a previous time step. If the supply of nodes in either A' or A'' is exhausted before all l_{pi} links have been allocated, then nodes in the other subset are linked to instead. The model is initialized by connecting the first node, p_i, to l_{pi} nodes in A'.

The exponential distribution used for λ_p and λ_{lp} is of the form $p(x) = \beta\exp(-\beta x)$, with $\beta = |P|(|A|-1)/(2(L-|P|)) - 1$. If we generalize our formalism and use a beta distribution[1,3], we do not find significantly different results (Supplementary Table 5).

Ratio of connectivity role norms. The ratio of the norms, d, is defined by $d = |x|/|y|$, where $|x| = (\sum_{i=1}^{m} x_i^2)^{1/2}$, $|y| = (\sum_{i=1}^{m} y_i^2)^{1/2}$, m is the number of connectivity roles and x_i and y_i are the numbers of nodes within each role i for the empirical and, respectively, model-generated networks. Ratios within $(0.9, 1.1)$ are fractions with a norm comparable to the empirical data.

Received 5 June; accepted 10 October 2008.
Published online 3 December 2008.

1. Williams, R. J. & Martinez, N. Simple rules yield complex food webs. *Nature* **404,** 180–183 (2000).
2. Cattin, M. F., Bersier, L. F., Banasek-Richter, C., Baltensperger, R. & Gabriel, J. P. Phylogenetic constraints and adaptation explain food-web structure. *Nature* **427,** 835–839 (2004).
3. Stouffer, D. B., Camacho, J., Guimerà, R., Ng, C. A. & Nunes Amaral, L. A. Quantitative patterns in the structure of model and empirical food webs. *Ecology* **86,** 1301–1311 (2005).
4. Dunne, J. A., Williams, R. J., Martinez, N. D., Wood, R. A. & Erwin, D. H. Compilation and network analyses of Cambrian food webs. *PLoS Biol.* **6,** 693–708 (2008).
5. Petchey, O. L., Beckerman, A. P., Riede, J. O. & Warren, P. H. Size, foraging, and food web structure. *Proc. Natl Acad. Sci. USA* **105,** 4191–4196 (2008).
6. Allesina, S., Alonso, D. & Pascual, M. A general model for food web structure. *Science* **320,** 658–661 (2008).
7. Bascompte, J. & Jordano, P. in *Ecological Networks: Linking Structure to Dynamics in Food Webs* (eds Pascual, M. & Dunne, J. A.) 143–159 (Oxford Univ. Press, 2006).
8. Jordano, P., Bascompte, J. & Olesen, J. M. Invariant properties in coevolutionary networks of plant-animal interactions. *Ecol. Lett.* **6,** 69–81 (2003).
9. Bascompte, J., Jordano, P., Melián, C. J. & Olesen, J. M. The nested assembly of plant–animal mutualistic networks. *Proc. Natl Acad. Sci. USA* **100,** 9383–9387 (2003).
10. Olesen, J. M., Bascompte, J., Dupont, Y. L. & Jordano, P. The modularity of pollination networks. *Proc. Natl Acad. Sci. USA* **104,** 19891–19896 (2007).
11. Sampson, R. J., Raudenbush, S. W. & Earls, F. Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science* **277,** 918–924 (1997).
12. Ostrom, E., Burger, J., Field, C. B., Norgaard, R. B. & Policansky, D. Revisiting the commons: Local lessons, global challenges. *Science* **284,** 278–282 (1999).
13. Hammerstein, P. (ed.) *Genetic and Cultural Evolution of Cooperation* (MIT Press, 2003).
14. Ohtsuki, H., Hauert, C., Lieberman, E. & Nowak, M. A. A simple rule for the evolution of cooperation on graphs and social networks. *Nature* **441,** 502–505 (2006).
15. Bronstein, J. L. The exploitation of mutualisms. *Ecol. Lett.* **4,** 277–287 (2001).
16. Waser, N. M., Chittka, L., Price, M. V., Williams, N. M. & Ollerton, J. Generalization in pollination systems, and why it matters. *Ecology* **77,** 1043–1060 (1996).
17. Noë, R. & Hammerstein, P. Biological markets: supply and demand determine the effect of partner choice in cooperation, mutualism and mating. *Behav. Ecol. Sociobiol.* **35,** 1–11 (1994).
18. Olesen, J. M. & Jordano, P. Geographic patterns in plant–pollinator mutualistic networks. *Ecology* **83,** 2416–2424 (2002).
19. Guimarães, P. R., Rico-Gray, V., Furtado dos Reis, S. & Thompson, J. N. Asymmetries in specialization in ant–plant mutualistic networks. *Proc. R. Soc. Lond. B* **273,** 2041–2047 (2006).
20. Rezende, E. L., Lavabre, J. E., Guimarães, P. R., Jordano, P. & Bascompte, J. Non-random coextinctions in phylogenetically structured mutualistic networks. *Nature* **448,** 925–928 (2007).
21. Santamaría, L. & Rodríguez-Gironés, A. Linkage rules for plant-pollinator networks: Trait complementarity or exploitation barriers? *PLoS Biol.* **5,** 354–362 (2007).
22. Guimarães, P. R. Jr *et al.* Building-up mechanisms determining the topology of mutualistic networks. *J. Theor. Biol.* **249,** 181–189 (2007).
23. Uzzi, B. The sources and consequences of embeddedness for the economic performance of organizations: the network effect. *Am. Sociol. Rev.* **61,** 674–698 (1996).
24. Carroll, G. R. & Hannan, M. T. *The Demography of Corporations and Industries* (Princeton Univ. Press, 2004).
25. Podolny, J. M. *Status Signals: A Sociological Study of Market Competition* (Princeton Univ. Press, 2005).
26. Amaral, L. A. N., Scala, A., Barthélémy, M. & Stanley, H. E. Classes of small-world networks. *Proc. Natl Acad. Sci. USA* **97,** 11149–11152 (2000).
27. Moody, J. & White, D. R. Structural cohesion and embeddedness: A hierarchical concept of social groups. *Am. Sociol. Rev.* **68,** 103–127 (2003).
28. Rodríguez-Gironés, M. A. & Santamaría, L. A new algorithm to calculate the nestedness temperature of presence–absence matrices. *J. Biogeography* **33,** 924–935 (2006).
29. Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. *Proc. Natl Acad. Sci. USA* **99,** 7821–7826 (2002).
30. Guimerà, R. & Nunes Amaral, L. A. Functional cartography of complex metabolic networks. *Nature* **433,** 895–900 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. Dunne, R. Guimerà, J. Kertész, M. Sales-Pardo, D. Stouffer and R. Williams for comments and suggestions. F.R.-T. acknowledges funding from the European Commission under the FP6 NEST Pathfinder Initiative 'Tackling Complexity in Science' (MMCOMNET project, contract no. 012999). S.S. held a Doctoral Research Studentship funded by MMCOMNET and CONACYT, and currently is supported by a Postdoctoral Fellowship at the Oxford University Corporate Reputation Centre in conjunction with the CABDyN Complexity Centre.

Author Contributions B.U. provided the NYGI data; F.R.-T. designed the research; S.S., F.R.-T. and B.U. analysed the data; S.S. ran the simulations; S.S. and F.R.-T. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to F.R-T. (felix.reed-tsochas@sbs.ox.ac.uk).

LETTERS

Ferritin is used for iron storage in bloom-forming marine pennate diatoms

Adrian Marchetti[1]*, Micaela S. Parker[1]*, Lauren P. Moccia[2], Ellen O. Lin[1], Angele L. Arrieta[3], Francois Ribalet[1], Michael E. P. Murphy[3], Maria T. Maldonado[2] & E. Virginia Armbrust[1]

Primary productivity in 30–40% of the world's oceans is limited by availability of the micronutrient iron[1,2]. Regions with chronically low iron concentrations are sporadically pulsed with new iron inputs by way of dust[3] or lateral advection from continental margins[4]. Addition of iron to surface waters in these areas induces massive phytoplankton blooms dominated primarily by pennate diatoms[5,6]. Here we provide evidence that the bloom-forming pennate diatoms *Pseudo-nitzschia* and *Fragilariopsis* use the iron-concentrating protein, ferritin, to safely store iron. Ferritin has not been reported previously in any member of the Stramenopiles, a diverse eukaryotic lineage that includes unicellular algae, macroalgae and plant parasites. Phylogenetic analyses suggest that ferritin may have arisen in this small subset of diatoms through a lateral gene transfer. The crystal structure and functional assays of recombinant ferritin derived from *Pseudo-nitzschia multiseries* reveal a maxi-ferritin that exhibits ferroxidase activity and binds iron. The protein is predicted to be targeted to the chloroplast to control the distribution and storage of iron for proper functioning of the photosynthetic machinery. Abundance of *Pseudo-nitzschia* ferritin transcripts is regulated by iron nutritional status, and is closely tied to the loss and recovery of photosynthetic competence. Enhanced iron storage with ferritin allows the oceanic diatom *Pseudo-nitzschia granii* to undergo several more cell divisions in the absence of iron than the comparably sized, oceanic centric diatom *Thalassiosira oceanica*. Ferritin in pennate diatoms probably contributes to their success in chronically low-iron regions that receive intermittent iron inputs, and provides an explanation for the importance of these organisms in regulating oceanic CO_2 over geological timescales[7,8].

Ferritin is an iron-storage protein used by plants, animals, cyanobacteria and other microorganisms to safely concentrate and store iron, thereby minimizing potential cell damage from reactive oxygen species and oxidative stress. Ferritin subunits from eukaryotes assemble into nanocages capable of storing up to 4,500 Fe(III) atoms as an iron oxide mineral by oxidizing iron at ferroxidase centres and releasing reduced iron on cellular demand[9]. We identified an expressed sequence tag (EST) for a gene encoding a ferritin-like protein in the pennate diatom *Pseudo-nitzschia australis*, and subsequently generated full-length sequences for this gene from *P. australis* and *P. multiseries* and partial sequences from additional species including oceanic *P. granii*. The predicted amino acid sequence of *Pseudo-nitzschia* ferritin displays general conservation of residues that form the ferroxidase centres; it also possesses the requisite signal peptide and plastid transit peptide motifs[10] to target ferritin subunits to the plastid (Supplementary Fig. 1), where they are expected to both store iron away from reactive oxygen as Fe(III) and

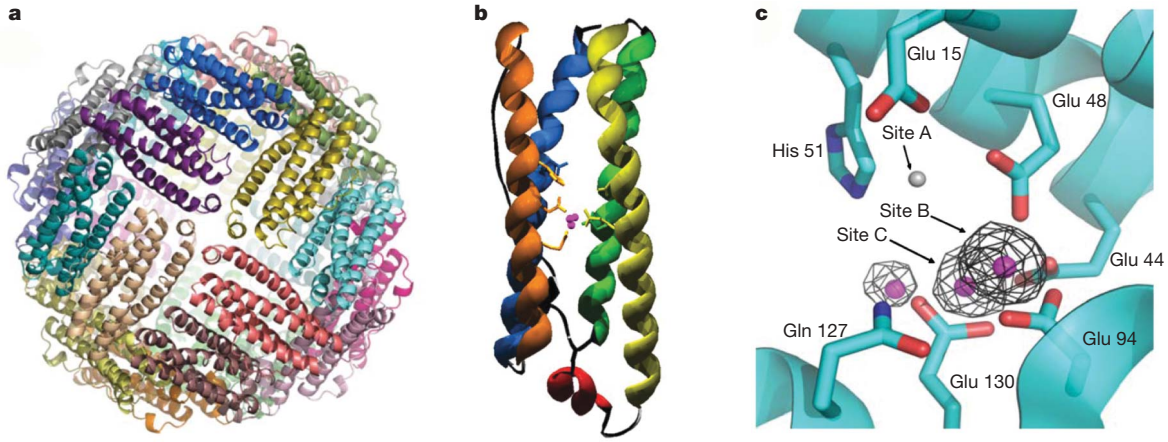


Figure 1 | Tertiary crystal structure and ferroxidase site of *Pseudo-nitzschia multiseries* ferritin. **a**, **b**, Crystal structure of the recombinant *P. multiseries* ferritin: **a**, multimer (24mer); **b**, monomer. α-helices and ferroxidase centre side chains: blue, A helix; orange, B helix; green, C helix; yellow, D helix; red, E helix. Iron atom (pink) positions are according to the crystal structure. **c**, Ferroxidase site A is occupied by a water molecule (grey sphere). One iron atom (pink sphere) occupies site B at full occupancy, and is coordinated by four glutamate residues (Glu 44, Glu 48, Glu 94 and Glu 130) shown as stick models. An iron atom at site C and a third iron atom are refined at occupancies of 50% and 40%, respectively. A difference anomalous dispersion map around the iron atoms is contoured at 3σ.

[1]School of Oceanography, University of Washington, Box 357940, Seattle, Washington 98195, USA. [2]Department of Earth and Ocean Sciences, University of British Columbia, 6270 University Boulevard, Vancouver, British Columbia V6T 1Z4, Canada. [3]Department of Microbiology and Immunology, University of British Columbia, 2350 Health Sciences Mall, Vancouver, British Columbia V6T 1Z3, Canada.
*These authors contributed equally to this work.

make Fe(II) available for iron-requiring processes, including photosynthesis. Residues that line the three-fold channels to facilitate iron release from the mineral core[11] are either conserved or conserved substitutions (Supplementary Fig. 1).

Characterization of recombinant *P. multiseries* ferritin (rPmFTN) was used to demonstrate that the protein displays features typical of maxi-ferritins. The apparent molecular mass of rPmFTN is ~530 kDa, consistent with formation of a 24-subunit structure (Supplementary Fig. 2). As measured by an aerobic oxidative iron uptake assay, rPmFTN displayed ferroxidase activity through the formation of oxy/hydroxo Fe(III) species (Supplementary Fig. 3a)[12]. Under these assay conditions, more than 600 iron atoms were loaded per apo-ferritin nanocage. The ferroxidase reaction consumed oxygen in a ratio of 1.9 ± 0.2 Fe(II):O_2 (Supplementary Fig. 3b), consistent with the proposed di-ferrous binding of oxygen at ferroxidase sites in maxi-ferritins[12,13]. Following the consumption of O_2, addition of catalase to the assay solution resulted in stoichiometric regeneration of O_2, indicating net production of H_2O_2 by the ferroxidase reaction as seen with many ferritins[12], but not bacterioferritins[14] or Dps ferritins[15].

The crystal structure of rPmFTN was resolved to 1.95 Å and provided further support that the *P. multiseries* protein functions as a ferritin *in vivo*. The structure revealed an assembly of 24 monomers with a typical maxi-ferritin arrangement (Fig. 1). The monomer fold is most similar (<1.3 Å root mean squared deviation, 155 aligned residues) to those of eukaryotic, bacterial and archeal ferritins, and less similar to bacterioferritins and Dps ferritins. Also, the ferroxidase site is conserved. Three iron atoms are observed in its vicinity; one iron atom is coordinated by conserved ferroxidase site B residues and a fourth ligand (Glu 44), and the other two iron atoms are positioned towards the core (Fig. 1c and Supplementary Fig. 4). Ferroxidase site A is occupied by a water molecule.

Iron nutritional status affected growth rates (μ) and ferritin transcript abundance for both coastal *P. multiseries* (North Atlantic) and oceanic *P. granii* (iron-limited northeast Pacific). *P. multiseries* had a reduced μ ($\mu/\mu_{max} = 0.2$; Fig. 2a) at a low total iron concentration, [Fe_T], of 6.2 nmol l^{-1} (see Supplementary Table 1 for other iron species' estimates). In contrast, *P. granii* maintained near-maximum μ ($\mu/\mu_{max} = 0.9$; Fig. 2a) at this [Fe_T], highlighting the significantly higher iron-use efficiencies of oceanic diatoms[16,17]. Under iron-limiting conditions, both species displayed lower μ, lower maximum photochemical yield of photosystem II (F_v/F_m), and lower chlorophyll *a* content (Fig. 2a and Supplementary Fig. 5a). For both species, the lowest amounts of ferritin transcripts corresponded to the lowest [Fe_T] (Fig. 2b). Significantly higher ferritin transcript abundances (one-way analysis of variance, Tukey test, $P < 0.05$) were observed at [Fe_T]

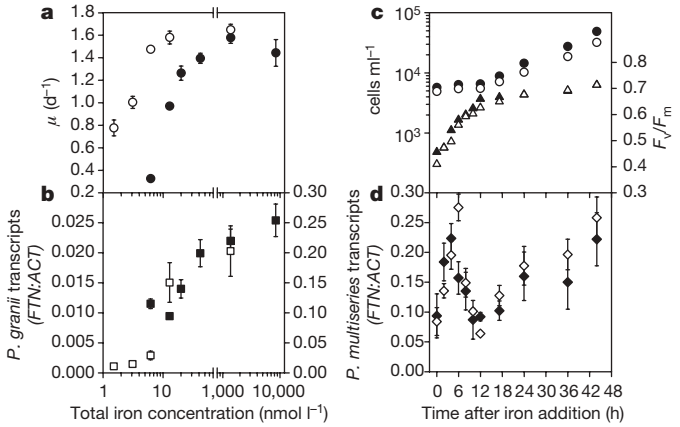


Figure 2 | Growth characteristics and ferritin transcript abundance in *Pseudo-nitzschia*. **a**, Growth rates (μ) of coastal *P. multiseries* (black filled circles) and oceanic *P. granii* (open circles) as a function of total iron concentration, [Fe_T]. **b**, Relative transcript abundance of ferritin normalized to actin (*FTN:ACT*) in *P. multiseries* (black filled squares) and *P. granii* (open squares) as a function of [Fe_T]. **c**, Maximum photochemical yield of photosystem II (F_v/F_m: black filled triangles, replicate A; open triangles, replicate B) and cell concentrations (black filled circles, replicate A; open circles, replicate B) as a function of time after 8 μmol l^{-1} iron addition to iron-limited *P. multiseries* originally grown at 6.2 nmol l^{-1} [Fe_T]. **d**, Relative transcript abundance of *FTN:ACT* (black filled diamonds, replicate A; open diamonds, replicate B) as a function of time after iron resupply. For **a** and **b**, symbols represent means of biological replicates ± s.e.m. ($n \geq 3$). For **d**, symbols represent means of technical triplicates ± s.d.

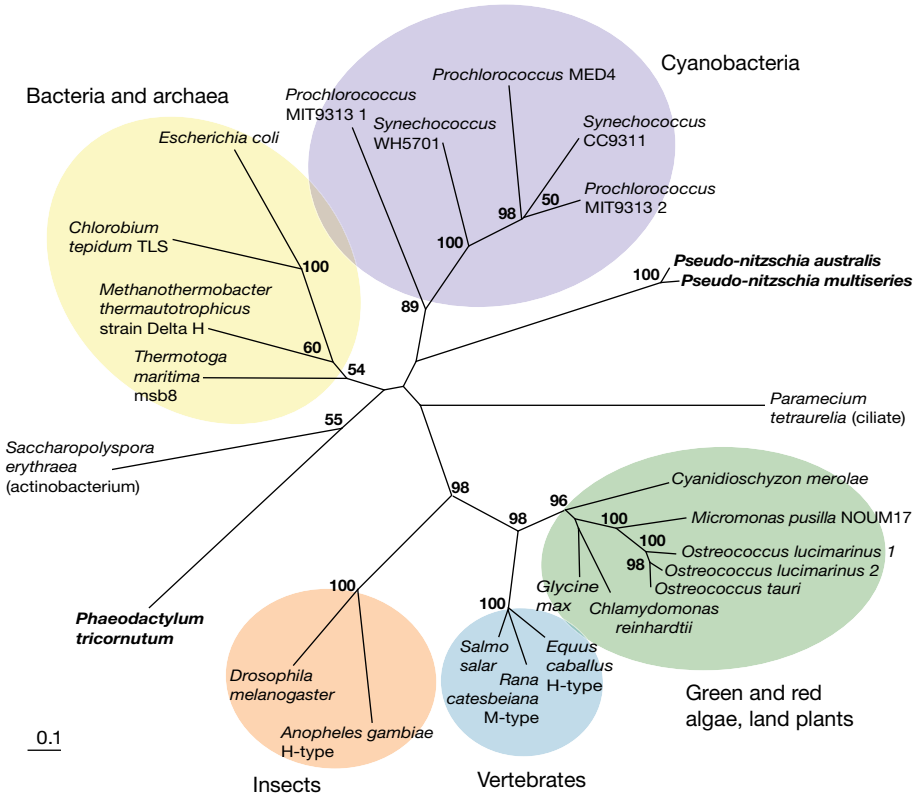


Figure 3 | Ferritin phylogenetic tree. Bootstrap consensus tree (100 replicates) showing the evolutionary relatedness of maxi-ferritins from 26 taxa inferred using a protein distance model (see Methods). Bootstrap values greater than 50 are indicated at the branch points. Diatoms are in bold.

that supported $\sim\mu_{max}$ (that is, $>42\,nmol\,l^{-1}$ for *P. multiseries*, and $>12.9\,nmol\,l^{-1}$ for *P. granii*). A $[Fe_T]$ of $12.9\,nmol\,l^{-1}$ is equivalent to estimates of artificial[18] and sporadic natural[4] iron inputs into iron-limited surface waters (see Supplementary Note 1). Over an identical range of $[Fe_T]$ ($6.2–1{,}370\,nmol\,l^{-1}$), oceanic *P. granii* displayed a 20-fold range in ferritin transcript abundance, whereas coastal *P. multiseries* displayed a twofold range, further emphasizing the enhanced sensitivity of oceanic species to ambient iron concentrations.

To mimic the sporadic availability of iron observed in nature, iron-limited *P. multiseries* cells were re-supplied with a pulse of iron. Coinciding with an immediate increase in F_v/F_m (Fig. 2c), ferritin transcript abundance increased and peaked within 6 h of iron addition, decreased to a minimum at 12 h and then gradually returned to levels measured in acclimated iron-replete cells (Fig. 2d). As in other eukaryotes[19], these results are consistent with *Pseudo-nitzschia* ferritin serving essential roles in both protection from oxidative stress (a rapid response to excess iron) and iron storage (a longer term acclimation).

Our identification of ferritin in pennate diatoms represents the first evidence of this iron-concentrating protein among the Stramenopiles, a major lineage of eukaryotes (Supplementary Fig. 6). We detected a ferritin-like gene in five additional open-ocean *Pseudo-nitzschia* species, in the closely-related Southern Ocean pennate diatom, *Fragilariopsis cylindrus* and in the whole genome sequence of the pennate diatom, *Phaeodactylum tricornutum*[20] (Supplementary Table 2 and Supplementary Fig. 1). Ferritin genes were not detected in available whole genome sequences of other Stramenopiles—*Thalassiosira pseudonana*[21] (centric diatom), *Aureococcus anophagefferens* (Pelagophyte) and two *Phytopthora* spp. (Oomycetes). Thus, there is no current evidence for iron storage by ferritin in centric diatoms or other Stramenopiles.

Ferritins from other photosynthetic eukaryotes (red and green algae, plants) are closely related to each other and to those found in animals (Fig. 3). In contrast, *P. multiseries* and *P. tricornutum* ferritins are only 24% identical (excluding predicted targeting peptides) and do not cluster with each other or with the other major maxi-ferritin branches, indicating that a common ancestor is not present among available ferritin sequences. The apparent absence of ferritin in the evolutionarily more ancient centric diatoms[22] in combination with large sequence disparities between *Phaeodactylum* and *Pseudo-nitzschia* suggest that the presence of ferritin in these diatoms may reflect lateral gene transfer events[23].

Consistent with their divergent evolutionary histories, pennate and centric diatoms have different iron acquisition[24] and storage strategies. Oceanic diatoms dramatically reduce their intracellular iron requirements under iron-limiting conditions[16,25,26]. Under high-iron conditions, however, oceanic *Pseudo-nitzschia* spp. store significantly more iron than oceanic *Thalassiosira* spp. (see Supplementary Note 2 and Supplementary Fig. 7). A continuous culture approach was implemented to determine whether iron storage by ferritin in oceanic *P. granii* provides a competitive advantage over the putative non-ferritin-containing oceanic centric diatom *T. oceanica*. In each experiment, cells were pre-acclimated to either high or low iron-replete conditions (μ_{max} maintained) before transfer to a continuous culture with iron-free medium (see Supplementary Note 3).

High-iron *P. granii* cells maintained >90% F_v/F_m and maximum cellular chlorophyll *a* fluorescence for 5.4 cell divisions and completed 9.2 cell divisions before μ decreased below 80% of μ_{max} (Fig. 4a and Supplementary Fig. 5b). Elevated ferritin transcript abundances were maintained for >4.5 cell divisions before rapidly declining to a minimum, suggesting a concomitant reduction in ferritin-stored iron. The decline in F_v/F_m as iron stores are depleted probably reflects a reduction in photosynthetic efficiency and reorganization of intracellular iron to maintain μ_{max}. Low-iron *P. granii* cells maintained >90% F_v/F_m for 1.6 divisions and >80% of μ_{max} for 2.8 divisions; ferritin transcript abundance began to decline almost immediately, consistent with minimal iron storage by these cells (Fig. 4b). Together, these results indicate that both F_v/F_m and ferritin transcript abundance are sensitive to ferritin iron stores and that this stored iron directly influences the number of cell divisions possible in the absence of added iron.

High-iron *T. oceanica* maintained >90% F_v/F_m for 3.1 cell divisions and >80% of μ_{max} for 5.3 cell divisions (Fig. 4c), indicating that iron stores in this diatom supports significantly fewer divisions than comparably grown *P. granii*. *P. granii* undergoes about 2.3 additional cell divisions before the decline in F_v/F_m and an additional 1.6 extra cell divisions before μ declines relative to *T. oceanica*, suggesting that approximately 60% ($2.3/(2.3 + 1.6) \times 100$) of the overall additional cell divisions by *P. granii* are due to enhanced iron storage. Assuming similar starting cell densities of *P. granii* and *T. oceanica*, 4 extra cell divisions by *P. granii* translates into ~16-fold more *P. granii* cells at the end of an iron-induced bloom. Enhanced iron storage in combination with higher iron-use efficiencies[16] explains why oceanic

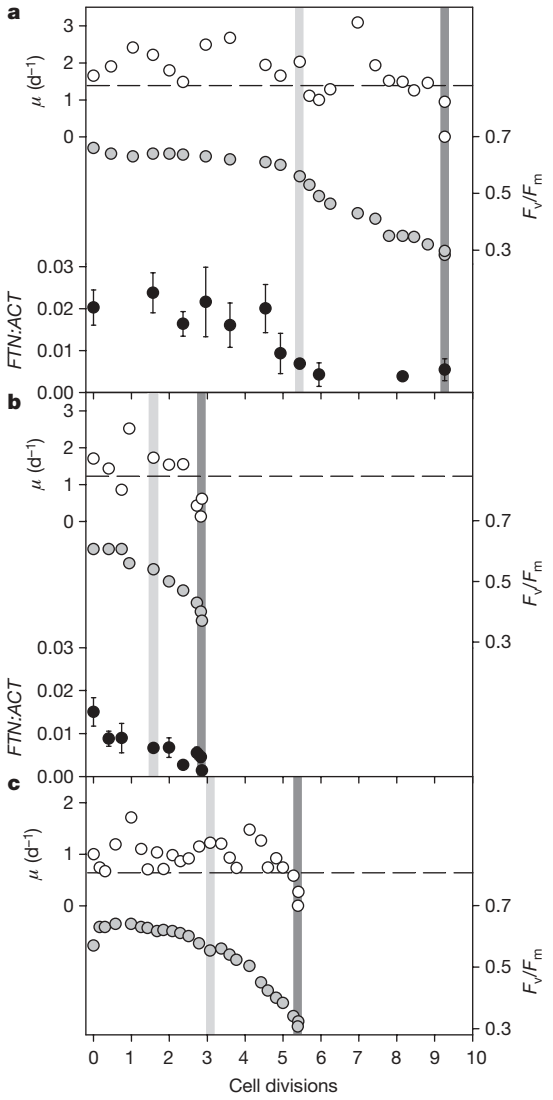


Figure 4 | Growth characteristics and ferritin transcript abundance in the oceanic diatoms *P. granii* and *T. oceanica* grown in iron-free continuous cultures. **a**, High iron ($1{,}370\,nmol\,l^{-1}$) pre-acclimated *P. granii*; **b**, low iron ($12.9\,nmol\,l^{-1}$) pre-acclimated *P. granii*; and **c**, high iron ($1{,}370\,nmol\,l^{-1}$) pre-acclimated *T. oceanica*. Shown are the 4-h interval growth rates (μ, open circles), maximum photochemical yield of photosystem II (F_v/F_m, grey filled circles) and relative transcript abundances of ferritin normalized to actin (*FTN:ACT*) (black filled circles; for *P. granii* only) as a function of cell divisions. Dashed lines indicate the dilution rate (set at 80% of μ_{max}). Light grey bars indicate the decline in F_v/F_m below 90% of the maximum ratio. Dark grey bars indicate the permanent decline in growth rates below the dilution rate. Error bars represent the s.d. of technical triplicates.

Pseudo-nitzschia commonly achieve cell densities an order of magnitude greater than oceanic centric species such as *T. oceanica* during iron-fertilization-induced blooms[6].

Identification of ferritin in a subset of pennate diatoms helps elucidate how some open ocean phytoplankton take advantage of pulsed iron supplies by safely sequestering large amounts of iron that support subsequent growth and divisions well after iron levels return to low, ambient concentrations. The ability to store iron may be crucial to achieving ample seed populations that persist between iron input events. The dominance of both *Pseudo-nitzschia* and *Fragilariopsis* during the numerous iron-enrichment experiments performed in iron-limited regions around the globe is probably not coincidental but rather due in part to high iron storage mediated through ferritin. The disparity between organismal and ferritin sequence phylogenies suggests that acquisition of ferritin may have been a relatively recent event that facilitated radiation of pennate diatoms into the open ocean (see Supplementary Note 4). Our results highlight how changes in the ocean environment, spanning near-instantaneous to geological timescales, shape the diversity and distributions of important groups of marine primary producers.

METHODS SUMMARY

Full-length *FTN* cDNA sequence was obtained using RACE. Cultures were grown using trace metal clean techniques in Aquil medium buffered with 100 $\mu mol\,l^{-1}$ of EDTA. Acclimated μ values were estimated from *in vivo* chlorophyll *a* fluorescence. Maximum photochemical yields of photosystem II were determined with a Phyto-PAM fluorometer (Walz). Cell counts were performed by microscopy using a Sedgwich-Rafter slide or with a Cytopeia Influx Cell Sorter flow cytometer. Total RNA (1–2 μg) was reverse transcribed using the 1st Strand cDNA Synthesis Kit with oligo-dT primers (Invitrogen) and used in qRT–PCRs with an iCycler iQ Real-Time PCR Detection System (Bio-Rad Laboratories) and linearized plasmids containing target genes as standards. Evolutionary relatedness of the *Pseudo-nitzschia* ferritin sequences to other ferritins was determined using the protein distance model JTT in the PHYLIP software package. The *P. multiseries FTN* genomic DNA lacking the signal peptide and plastid targeting sequences was overexpressed in *Escherichia coli* and rPmFTN was purified by anion exchange chromatography. Purity and molecular weight of rPmFTN monomer were determined by SDS–PAGE and the size of purified rPmFTN was determined with a calibrated gel filtration column. Ferroxidase activity was assayed spectroscopically by monitoring the increase in absorbance during aerobic titration of a 0.76 $\mu mol\,l^{-1}$ solution of purified rPmFTN (in 50 $mmol\,l^{-1}$ MES, 100 $mmol\,l^{-1}$ NaCl, pH 6.5) with ferrous sulphate solution in ~8 Fe equivalents up to ~1,000 Fe(II):ferritin. Oxygen consumption was monitored using an Apollo 4000 Free Radical Analyser fitted with a Clark type electrode. Crystals of rPmFTN were obtained from 1.4–1.6 $mol\,l^{-1}$ ammonium sulphate and 0.1 $mol\,l^{-1}$ ammonium acetate buffer pH 5.5. X-ray data were collected from iron soaked and apo-ferritin crystals and the structure was solved by molecular replacement using the *Pyrococcus furiosus* ferritin (PDB ID 2JD6) as a model. Anomalous dispersion data were used to identify iron atoms in the structure.

Received 22 January; accepted 13 October 2008.
Published online 26 November 2008.

1. Martin, J. H. & Fitzwater, S. Iron deficiency limits phytoplankton growth in the north-east Pacific subarctic. *Nature* **331,** 341–343 (1988).
2. Moore, J. K., Doney, S. C., Glover, D. M. & Fung, I. Y. Iron cycling and nutrient-limitation patterns in surface waters of the World Ocean. *Deep-Sea Res. II* **49,** 463–507 (2002).
3. Jickells, T. D. *et al.* Global iron connections between desert dust, ocean biogeochemistry, and climate. *Science* **308,** 67–71 (2005).
4. Lam, P. J. & Bishop, J. K. B. The continental margin is a key source of iron to the HNLC North Pacific Ocean. *Geophys. Res. Lett.* **35,** L07608, doi:10.1029/2008GL033294 (2008).
5. Marchetti, A., Sherry, N. D., Kiyosawa, H., Tsuda, A. & Harrison, P. J. Phytoplankton processes during a mesoscale iron enrichment in the NE subarctic Pacific: Part I – biomass and assemblage. *Deep-Sea Res. II* **53,** 2095–2113 (2006).
6. de Baar, H. J. W. *et al.* Synthesis of iron fertilization experiments: From the iron age in the age of enlightenment. *J. Geophys. Res.* **110,** C09S16, doi:10.1029/2004JC002601 (2005).
7. Cortese, G. & Gersonde, R. Morphometric variability in the diatom *Fragilariopsis kerguelensis*: Implications for Southern Ocean paleoceanography. *Earth Planet. Sci. Lett.* **257,** 526–544 (2007).
8. Katsuki, K. & Takahashi, K. Diatoms as paleoenvironmental proxies for seasonal productivity, sea-ice and surface circulation in the Bering Sea during the late Quaternary. *Deep-Sea Res. II* **52,** 2110–2130 (2005).
9. Liu, X. & Theil, E. C. Ferritins: Dynamic management of biological iron and oxygen chemistry. *Acc. Chem. Res.* **38,** 167–175 (2005).
10. Gruber, A. *et al.* Protein targeting into complex diatom plastids: functional characterisation of a specific targeting motif. *Plant Mol. Biol.* **64,** 519–530 (2007).
11. Jin, W., Takagi, H., Pancorbo, B. & Theil, E. C. "Opening" the ferritin pore for iron release by mutation of conserved amino acids at interhelix and loop sites. *Biochemistry* **40,** 7525–7532 (2001).
12. Yang, X. O., Chen-Barret, Y., Arosio, P. & Chasteen, N. D. Reaction paths of iron oxidation and hydrolysis in horse spleen and recombinant human ferritins. *Biochemistry* **37,** 9743–9750 (1998).
13. Xu, B. & Chasteen, N. D. Iron-oxidation chemistry in ferritin. *J. Biol. Chem.* **266,** 19965–19970 (1991).
14. Bou-Abdallah, F., Lewin, A. C., Le Brun, N. E., Moore, G. R. & Chasteen, N. D. Iron detoxification properties of *Escherichia coli* bacterioferritin — Attenuation of oxyradical chemistry. *J. Biol. Chem.* **277,** 37064–37069 (2002).
15. Yang, X., Chiacone, E., Stefanini, S., Ilari, A. & Chasteen, N. D. Iron oxidation and hydrolysis reactions of a novel ferritin from *Listeria innocua. Biochem. J.* **349,** 783–786 (2000).
16. Marchetti, A., Maldonado, M. T., Lane, E. S. & Harrison, P. J. Iron requirements of the pennate diatom *Pseudo-nitzschia*: Comparison of oceanic (HNLC) and coastal species. *Limnol. Oceanogr.* **51,** 2092–2101 (2006).
17. Sunda, W. G., Swift, D. G. & Huntsman, S. A. Low iron requirement for growth in oceanic phytoplankton. *Nature* **351,** 55–57 (1991).
18. Bowie, A. R. *et al.* The fate of added iron during a mesoscale fertilisation experiment in the Southern Ocean. *Deep-Sea Res. II* **48,** 2703–2743 (2001).
19. La Fontaine, S. *et al.* Copper-dependent iron assimilation pathway in the model photosynthetic eukaryote *Chlamydomonas reinhardtii. Eukaryot. Cell* **1,** 736–757 (2002).
20. Phaeodactylum. *tricornutum* v2.0. ⟨http://genome.jgi-psf.org/Phatr2/Phatr2.home.html⟩ (2006).
21. Armbrust, E. V. *et al.* The genome of the diatom *Thalassiosira pseudonana*: Ecology, evolution, and metabolism. *Science* **306,** 79–86 (2004).
22. Sims, P. A., Mann, D. G. & Medlin, L. K. Evolution of the diatoms: Insights from fossil, biological and molecular data. *Phycologia* **45,** 361–402 (2006).
23. Andersson, J. O. Lateral gene transfer in eukaryotes. *Cell. Mol. Life Sci.* **62,** 1182–1197 (2005).
24. Kustka, A. B., Allen, A. E. & Morel, F. M. M. Sequence analysis and transcriptional regulation of iron acquisition genes in two marine diatoms. *J. Phycol.* **43,** 715–729 (2007).
25. Maldonado, M. T. & Price, N. M. Influence of N substrate on Fe requirements of marine centric diatoms. *Mar. Ecol. Prog. Ser.* **141,** 161–172 (1996).
26. Sunda, W. & Huntsman, S. A. Iron uptake and growth limitation in oceanic and coastal phytoplankton. *Mar. Chem.* **50,** 189–206 (1995).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank S. S. Bates for providing *P. multiseries* CLN-47 and CLN-17 isolates used in this study, and T. Rynearson for sea water from OSP in which *P. granii* (UWOSP36) was isolated. We are grateful to K. R. Boissonneault for providing *P. multiseries* actin sequences, R. Marohl and I. Oleinkov for assistance with culture maintenance and sample analyses, and W. C. Lee for assistance with X-ray diffraction data collection. This study was supported by a Gordon and Betty Moore Foundation Marine Microbiology Investigator Award, National Science Foundation grants and a National Institute of Environmental Health Sciences grant to E.V.A.; a National Sciences and Engineering Research Council of Canada grant to M.T.M.; and a Canadian Institutes of Health Research grant to M.E.P.M. Portions of this research were carried out at the Canadian Light Source (CLS) and the Stanford Synchrotron Radiation Laboratory (SSRL). SSRL is a national user facility operated by Stanford University on behalf of the US Department of Energy, Office of Basic Energy Sciences. The SSRL Structural Molecular Biology Program is supported by the Department of Energy, Office of Biological and Environmental Research, and by the National Institutes of Health, National Center for Research Resources, Biomedical Technology Program, and the National Institute of General Medical Sciences.

Author Contributions A.M., M.S.P. and E.V.A. designed the study; A.M. and M.S.P. performed the *FTN* annotation, phylogeny analyses and conducted the *FTN* expression and continuous culture experiments; M.E.P.M. and M.T.M. guided—and L.P.M. and A.A. conducted—the ferritin protein crystallography and biochemical characterization experiments; E.O.L. assisted with sequencing and performed *FTN* surveys; F.R. performed flow cytometry analyses and assisted with culture experiments; A.M. wrote the paper with assistance from E.V.A., M.S.P., L.P.M., M.T.M., A.A. and M.E.P.M. All authors discussed the results and commented on the manuscript.

Author Information All sequences have been deposited in GenBank under accession codes FJ004953–FJ004969 and are listed in Supplementary Table 2. The coordinates of the apo and iron-soaked ferritin structures have been deposited in the Protein Data Bank under respective accession numbers 3E6R and 3E6S. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to E.V.A. (armbrust@ocean.washington.edu).

LETTERS

Discovery of a sexual cycle in the opportunistic fungal pathogen *Aspergillus fumigatus*

Céline M. O'Gorman[1,2], Hubert T. Fuller[1] & Paul S. Dyer[2]

***Aspergillus fumigatus* is a saprotrophic fungus whose spores are ubiquitous in the atmosphere[1]. It is also an opportunistic human pathogen in immunocompromised individuals, causing potentially lethal invasive infections[2,3], and is associated with severe asthma and sinusitis[4]. The species is only known to reproduce by asexual means[5], but there has been accumulating evidence for recombination and gene flow from population genetic studies[5–8], genome analysis[9,10], the presence of mating-type genes[8,10] and expression of sex-related genes[8] in the fungus. Here we show that *A. fumigatus* possesses a fully functional sexual reproductive cycle that leads to the production of cleistothecia and ascospores, and the teleomorph *Neosartorya fumigata* is described. The species has a heterothallic breeding system; isolates of complementary mating types are required for sex to occur. We demonstrate increased genotypic variation resulting from recombination between mating type and DNA fingerprint markers in ascospore progeny from an Irish environmental subpopulation. The ability of *A. fumigatus* to engage in sexual reproduction is highly significant in understanding the biology and evolution of the species. The presence of a sexual cycle provides an invaluable tool for classical genetic analyses and will facilitate research into the genetic basis of pathogenicity and fungicide resistance in *A. fumigatus*, with the aim of improving methods for the control of aspergillosis. These results also yield insights into the potential for sexual reproduction in other supposedly 'asexual' fungi.**

A. fumigatus is found worldwide in soils and on organic debris, where it is important in nutrient recycling[1,2]. The species has long been considered asexual, with dispersal achieved by the abundant production of haploid conidia[5]. Inhaled conidia are normally eliminated by the innate immune response. However, *A. fumigatus* has become the most prevalent airborne fungal pathogen as a result of its ability to cause infections in immunocompromised hosts, with a human mortality rate of at least 50% (refs 2, 3, 11). Airborne conidia are also allergens, associated with asthma, allergic sinusitis and bronchoalveolitis[2,4].

The importance of *A. fumigatus* to human health has led to genome sequencing and investigations into the population biology of the species[7,9,12]. Studies have revealed that, despite its supposedly asexual status, *A. fumigatus* has many characteristics of a sexual species. Genome screening detected 215 genes implicated in sexual development[10], including a high-mobility group (HMG)-domain gene located at a 'mating-type' (*MAT*) locus typical of sexually reproducing heterothallic (obligately outcrossing) euascomycete fungi[13]. Such HMG genes are believed to be ancestral sex determinants in fungi[14]. For sex to occur in heterothallic fungi, it is necessary for isolates of opposite *MAT1-1* and *MAT1-2* mating types to be present[15]; by convention these contain either a *MAT1-1* alpha-domain or *MAT1-2* HMG-domain gene, respectively, at the *MAT* locus[13]. These *MAT* genes encode transcription factors that are master regulators of sexual reproduction, required for normal sexual development in *Aspergillus* and other species[13,16]. A related study[8] identified isolates of *A. fumigatus* containing the complementary alpha-domain *MAT* gene, and showed expression of *MAT1-1* and *MAT1-2* and of genes encoding sex pheromones and pheromone receptors. In addition, a nearly 1:1 ratio of *MAT1-1*:*MAT1-2* isolates was found among a worldwide screening of 290 isolates, which is consistent with latent sexuality[8]. Gene recombination was also detected within a subset of isolates[8], in agreement with other studies of *A. fumigatus* showing evidence of recombination and no detectable population structure, albeit with predominantly clonal reproduction[6,7,12,17]. The observed recombination was attributed to past meiotic events, a so far undetected 'cryptic' sexual state or parasexual gene flow[5,6,8].

Previous population studies of *A. fumigatus* involved analysis of global collections or limited regional sample sets[7,8,12,17]. In this study we screened for mating type[8] in an Irish environmental population of *A. fumigatus*, composed of 91 isolates collected from five locations in Dublin during 2005 in the course of aerial sampling work[18] (Supplementary Table 1). Isolates were also fingerprinted by randomly amplified polymorphic DNA (RAPD)-PCR to detect clonality and assess genetic relatedness (see Methods). *MAT1-1* or *MAT1-2* specific bands amplified from all except one isolate (Supplementary Fig. 1). Eighty-eight of the isolates had unique RAPD profiles, showing genetic variability within the Dublin population. A ratio of 49.4% *MAT1-1* to 50.6% *MAT1-2* was evident in the clone-corrected data set, with no significant difference in mating-type distribution at any of the collection sites ($P > 0.05$; Supplementary Table 2) and no evidence for clustering of isolates according to collection site or mating type (Fig. 1). Indeed, both mating types were found in samples collected within minutes of each other, showing that *MAT1-1* and *MAT1-2* isolates are in close proximity in nature. Results are in agreement with previous studies that found *MAT* genotypes in nearly equal proportions[8,12], although one survey reported a predominance of *MAT1-2* (ref. 17).

Species identity of the Dublin isolates was verified by molecular analysis, given possible misidentification of *A. fumigatus sensu stricto* by morphotyping[19,20]. A subset of 12 representative isolates (6 *MAT1-1*, 6 *MAT1-2*) was chosen (Supplementary Table 1), and regions of their β-tubulin (*benA*) and carboxypeptidase-5 (*cxp*) genes were sequenced (S. A. Balajee, D. Nickle, L. Razai, S. F. Hurst & K. A. Marr, personal communication; see Methods). All isolates had identical sequences to those of known *A. fumigatus* depositions (for example, GenBank AY685162 (*benA*) and DQ438742 (*cxp*)) except for AfIR931, which had only one nucleotide difference in the *cxp* amplicon.

We then set up crosses with the 12 confirmed *A. fumigatus sensu stricto* isolates in all possible combinations of opposite mating types ($n = 36$) on a range of growth media at different temperatures to

[1]UCD School of Biology and Environmental Science, University College Dublin, Belfield, Dublin 4, Ireland. [2]School of Biology, University of Nottingham, University Park, Nottingham NG7 2RD, UK.

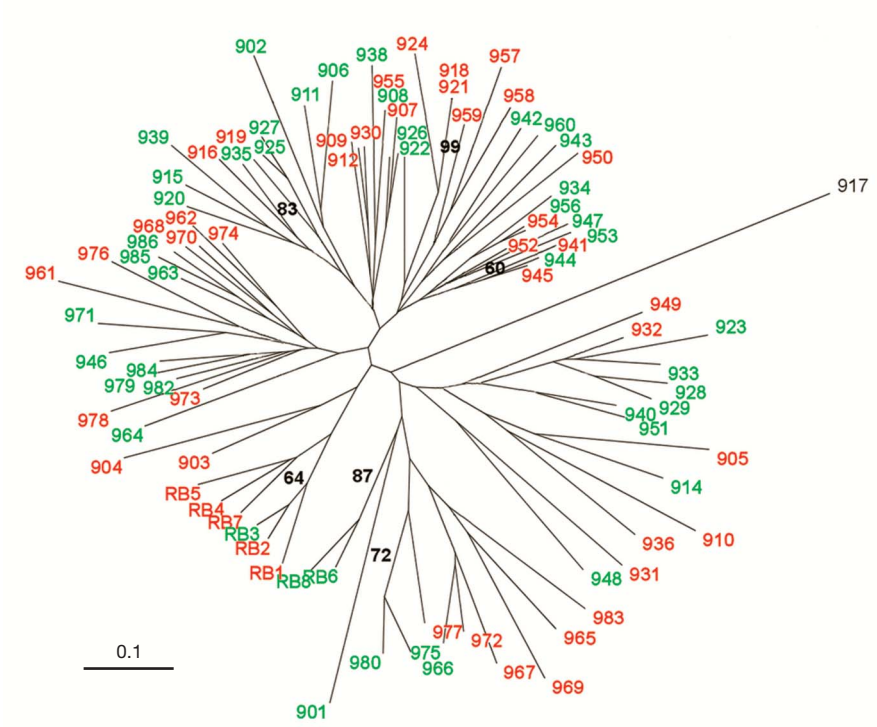


Figure 1 | Genetic diversity of 91 *A. fumigatus* isolates from Dublin, Ireland. Radial dendrogram derived from RAPD-PCR data pooled from primers UBC90, R151, R108 and RC08. The tree was constructed by neighbour-joining analysis with branch lengths drawn to reflect genetic distance derived from Jaccard's coefficient of RAPD band matching (scale bar: 0.1 = 10% genetic difference). Numbers at nodes indicate the percentage bootstrap support (based on 1,000 resamplings) for which the cluster is supported (only values of 60% or more are shown, for clarity). *MAT1-1* and *MAT1-2* genotypes are shown in red and green, respectively.

determine whether it was possible to induce sexual reproduction *in vitro* (see Methods). After 6 months of incubation, mature cleistothecia (spherical fruiting bodies characteristic of sex in the aspergilli[21]) were found on pairings grown on Parafilm-sealed Oatmeal agar[22] plates at 30 °C in the dark (Fig. 2a–d and Table 1). The light-yellow cleistothecia (typically 150–500 μm in diameter) formed singly or in small clusters of two to five, mainly along the junction (barrage zone) where hyphae of the parental isolates came into contact, and to a smaller extent within mycelium on either side of the barrage zone. The related heterothallic species *Neosartorya spathulata* produces cleistothecia scattered or grouped in clusters[23], whereas *N. fennelliae* generally produces copious numbers of cleistothecia along lines of mycelial contact[24]. There was marked variation in the numbers of cleistothecia produced between different pairings (Table 1). It is not yet known whether this variation is typical of the species. Cleistothecia failed to develop when other media and incubation temperatures were used, or in pairings between isolates of the same mating type. When cleistothecia were squashed, numerous yellowish-white to greenish-white lenticular ascospores (4–5 μm in diameter) with two equatorial crests were observed, together with occasional intact asci (Fig. 2e–g). The ridged ornamentation of the ascospores is similar to that in the homothallic species *N. assulata*[20]. The teleomorph (sexual stage) of *A. fumigatus* was assigned to the genus *Neosartorya* on the basis of phylogenetic relatedness and morphology of the cleistothecia and ascospores[12,20], and named *Neosartorya fumigata* O'Gorman, Fuller & Dyer *sp. nov.*

Ascospores of *N. fumigata* germinated on 2% malt extract agar (MEA) at 28 °C within 15 h and also germinated after exposure to 70 °C for 90 min (Fig. 2h), heat resistance of ascospores being typical of *Neosartorya* species[20]. Ascospore-derived cultures gave rise to characteristic *A. fumigatus* colonies bearing abundant asexual conidia under routine growth conditions.

Finally, 15 ascospore progeny from each of three different crosses were assessed for meiotic recombination by examining the segregation of five genetic markers (four RAPD bands and the mating-type genotype) (Fig. 3). Between 53% and 67% of progeny showed unique genotypes, with only 7–13% of progeny identical to one of the parents (on the basis of the markers examined), providing clear evidence of recombination during the sexual cycle (Supplementary Tables 3–5). The segregation of mating type was consistent with a heterothallic breeding system. Fisher's exact test confirmed that genetic markers were equally distributed between the progeny in each cross, verifying the null hypothesis of 1:1 Mendelian segregation as a result of independent assortment.

A. fumigatus was first described 145 years ago[20], but its teleomorph has remained undiscovered until now. Indeed, previous efforts failed to induce sexual reproduction despite considerable effort[25] (G. S. May, N. Osherov and J. Kwon-Chung, personal communication). There are several possible explanations of why the sexual cycle has not been reported before. Environmental parameters rarely encountered in nature might be required to trigger sex[15], or cleistothecial production may be confined to substrata that have not been systematically monitored. Sexuality has now been demonstrated in the once presumed asexual pathogen *Candida albicans*, with mating suggested to occur only in particular host niches[26]. Similarly, sexual reproduction has been detected in previously 'asexual' phytopathogenic *Tapesia* and *Mycosphaerella* species on certain host substrates[27,28]. Alternatively the species as a whole may be experiencing a 'slow decline' in fertility[5] — possibly linked to variation in *MAT* gene expression[5,25]. Thus, the Dublin isolates might represent a rare subset of fertile isolates, although there is evidence that *A. fumigatus* comprises one global, recombining population[7,12] (see Supplementary Discussion).

The discovery of a sexual cycle in *A. fumigatus* provides insights into the biology and evolution of the species. It helps explain the presence of diverse genotypes despite predominantly clonal reproduction[6,7,17], conservation of sex-related genes[10,29], aspects of genome evolution[29] and defence against repetitive elements[2]. In addition, ascospores might aid survival in adverse environmental conditions. The discovery also has significant medical implications. Sexual

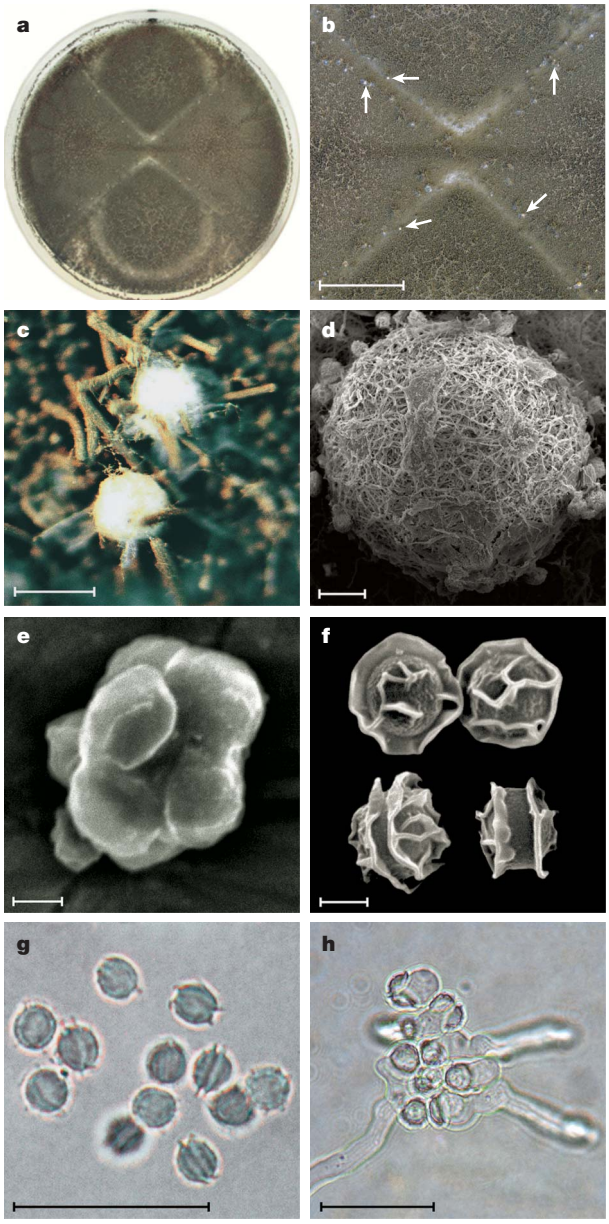


Figure 2 | *Neosartorya fumigata sp. nov.* **a**, Paired culture of AfRB2 × AfIR956 incubated for 6 months at 30 °C on Oatmeal agar medium in a 9-cm diameter Petri dish. **b**, Cleistothecia (arrows) along junctions of intersecting colonies of opposite mating type. Scale bar, 1 cm. **c**, Cleistothecia among chains of conidia. Scale bar, 400 μm. **d**, Scanning electron micrograph of a cleistothecium, showing the peridium of interwoven hyphae. Scale bar, 100 μm. **e**, **f**, Scanning electron micrograph of an eight-spored ascus (**e**) and ascospores (**f**). Scale bars, 2 μm. **g**, **h**, Ascospores ungerminated (**g**) and germinating after exposure to 70 °C for 60 min (**h**). Scale bars, 20 μm.

Table 1 | Numbers of cleistothecia produced by *A. fumigatus* crosses

Cross		Number of cleistothecia					
		MAT1-1					
		AfIR909	AfIR931	AfIR954	AfIR957	AfIR974	AfRB2
MAT1-2	AfIR901	+	+	+	−	+++++	+
	AfIR915	+	+	−	+	+++	++
	AfIR928	++	>	+++	++	++++	>
	AfIR956	+	++	+	+	++	++
	AfIR964	++	+	+	++++	>	++
	AfRB3	+	+	+	++	+++	+++++

Ratings indicate the mean number of cleistothecia produced from two replicate crosses on Oatmeal agar medium in 9-cm diameter Petri dishes after incubation in the dark for 6 months at 30 °C: −, none; +, 1–19; ++, 20–39; +++, 40–59; ++++, 60–79; +++++, 80–100; >, more than 100 cleistothecia.

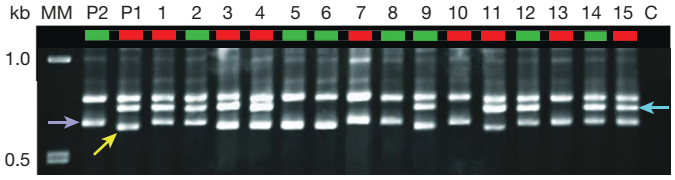


Figure 3 | Segregation patterns of molecular markers. Representative RAPD-PCR data from *N. fumigata* parental isolates (P1 and P2) and 15 ascospore progeny from the cross AfIR974 × AfIR964 using primer OPW-10. MM, molecular mass marker; P2, AfIR964; P1, AfIR974; lanes 1–15, progeny 974-964-1, 974-964-3 to 974-964-16; C, water control. Coloured arrows indicate bands designated OPW-10-1 (purple), OPW-10-2 (yellow) and OPW-10-3 (blue) (Supplementary Table 3). Lane headings are shown in red and green to indicate *MAT1-1* and *MAT1-2* genotypes, respectively. kb, kilobase.

reproduction can result in progeny with increased virulence or resistance to antifungal agents, and can confound diagnostic tests based on the assumption of clonality[5,30]. However, the sexual cycle also offers a valuable tool with which to determine the genetic basis of traits of interest. Pathogenicity in *A. fumigatus* seems multifactorial in nature[2,11], with genome analysis revealing potential targets[9]. Classical genetic analyses may help in elucidating the contribution of particular components, thereby guiding disease control strategies.

Results are also of general significance, given that almost one-fifth of all fungi have no known sexual stage. This includes many *Aspergillus*, *Penicillium*, *Coccidioides* and *Malassezia* species of great economic and medical importance. Some of these species have apparently functional *MAT* and other sex-related genes[10,21] (Supplementary Discussion). Recombination in asexual populations is often attributed to parasexual reproduction. However, parasexuality is limited to isolates of the same vegetative compatibility grouping[15]. According to the *A. fumigatus* model, cryptic sexuality is a more likely source of recombination, and isolation of fresh cultures combined with concerted laboratory crossing efforts of compatible mating types may lead to a sexual revolution for many of these other supposed 'asexuals'.

Neosartorya fumigata O'Gorman, Fuller & Dyer sp. nov.

Etymology. The name *fumigata* is derived from the Latin *fumigare* 'to make smoke', which refers to the clouds of spores produced by the fungus.

Holotype. Dried oatmeal agar plate with teleomorph of paired *Aspergillus fumigatus* colonies, AfRB2 (*MAT1-1*) × AfIR956 (*MAT1-2*), both collected by C. O'Gorman from outdoor air at Belfield, Dublin, November and December 2005, respectively; deposited in the Herbarium of the Royal Botanic Gardens, Kew (K) (K(M)159484). Isotypes in K (K(M)159485) and the National Herbarium, National Botanic Gardens, Dublin (DBN), each with a slide preparation of squashed cleistothecia.

Referred material. Paired cultures of *A. fumigatus* AfIR957 (*MAT1-1*) × AfIR928 (*MAT1-2*) and of AfIR974 (*MAT1-1*) × AfIR964 (*MAT1-2*) (see Supplementary Table 1 for isolate details).

Latin diagnosis. *Fungus heterothallicus. Cleistothecia superficialia, cincta gossypio mycelio, globosa, 150–480 (600) μm in diametro, dispersa singula vel coniuncta, primo alba vel luteo-alba, deinde mutans ad dilutum luteam vel flavo-griseam. Peridium factum est ex permultis propaginibus anguste intertextis, et interdum amplificatis hyphis, in tempore cretum simule membranae. Asci varie subglobosi, octospori, maturi evanescentes. Ascosporae luteae albae ad pallidas virentes, unicellulares, late lenticulares, raro globosae, cum duabus cristis aequatorialibus—0.3–1.0 μm latis, sine cristis 4–5 μm; superficies convexae cum costis reticulatis proiectis, reticulum variat, ac costae effusae usque ad cristas; levis inter cristas, aliquando lapilli in medio earum. Status anamorphus: Aspergillus fumigatus* Fresenius.

Diagnosis. Heterothallic fungus. Cleistothecia superficial, surrounded by cottony mycelium, globose, 150–480 (600) μm in diameter, scattered or grouped in small clusters, at first white to yellowish white

(4A2) then turning light yellow (4A4) to greyish yellow (4B5). Peridium composed of several layers of tightly interwoven, occasionally flattened hyphae; membranous with age. Asci irregularly subglobose, eight-spored, evanescent at maturity. Ascospores yellowish white (1A2) to greenish white (28A2), one-celled, broadly lenticular, rarely spherical, with two equatorial crests—0.3–1.0 μm wide, spore body 4–5 μm; convex surface with prominent ridges forming a reticulate ornamentation, degree of reticulation varies, ridges spreading to the equatorial crest; area between crests generally smooth, occasionally with a midline of small projections. Anamorph: *Aspergillus fumigatus* Fresenius.

METHODS SUMMARY

Aspergillus fumigatus isolates were crossed in pairwise *MAT1-1* and *MAT1-2* combinations in duplicate on Oatmeal agar medium[22] (Pinhead Oatmeal; Odlum Group) and were incubated inverted at 15, 30 or 45 °C in the dark. Crosses were also established on 2% MEA (Oxoid), Czapek Dox agar and *Aspergillus* complete medium[8] at 30 °C in the dark. Spore suspensions of each isolate (5×10^5 conidia ml^{-1}) were prepared from seven-day-old cultures. Two 1-μl aliquots of each spore suspension were separately inoculated onto the agar surface about 4 cm apart and perpendicular to aliquots of conidia of the opposite mating type. This configuration created four interaction/barrage zones as colonies grew (see Fig. 2a, b). Plates were sealed with one layer of Parafilm. Crosses were examined for cleistothecia periodically over 6 months with Olympus SZH10 Stereo and BX45 light microscopes. Cleistothecia were removed with a sterile needle tip and cleaned of adhering conidia by gentle rolling across a drop of water under sterile conditions. Individual cleistothecia were added to 100 μl of 0.05% Tween 80 and ruptured by squashing with a needle tip. The resulting macerate (of peridium, asci, ascospores and contaminating conidia) was transferred to 1.9 ml of 0.05% Tween 80 and vortex-mixed for 1 min to disrupt the asci. The suspension was heated at 70 °C for 60 min, which prevented the germination of conidia. Aliquots (100 μl) of the suspension were spread plated on 2% MEA (9-cm plates) and incubated overnight at 28 °C. Single-ascospore cultures were established on 2% MEA by transferring individual germinating ascospores with a LaRue lens cutter attached to an Olympus BX45 microscope.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 13 August; accepted 9 October 2008.
Published online 30 November 2008; corrected 22 January 2009 (details online).

1. Mullins, J., Harvey, R. & Seaton, A. Sources and incidence of airborne *Aspergillus fumigatus* (Fres). *Clin. Allergy* **6,** 209–217 (1976).
2. Latgé, J. P. *Aspergillus fumigatus* and aspergillosis. *Clin. Microbiol. Rev.* **12,** 310–350 (1999).
3. Lin, S. J., Schranz, J. & Teutsch, S. M. Aspergillosis case-fatality rate: systematic review of the literature. *Clin. Infect. Dis.* **32,** 358–366 (2001).
4. Anderson, M. J., Brookman, J. L. & Denning, D. W. in *Genomics of Plants and Fungi* (eds Prade, R.A. & Bohnert, B.J.) 1–39 (Marcel Dekker, 2003).
5. Dyer, P. S. & Paoletti, M. Reproduction in *Aspergillus fumigatus*: sexuality in a supposedly asexual species? *Med. Mycol.* **43** (Suppl. 1), 7–14 (2005).
6. Varga, J. & Tóth, B. Genetic variability and reproductive mode of *Aspergillus fumigatus*. *Infect. Genet. Evol.* **3,** 3–17 (2003).
7. Pringle, A. *et al.* Cryptic speciation in the cosmopolitan and clonal human pathogenic fungus *Aspergillus fumigatus*. *Evolution* **59,** 1886–1899 (2005).
8. Paoletti, M. *et al.* Evidence for sexuality in the opportunistic human pathogen *Aspergillus fumigatus*. *Curr. Biol.* **15,** 1242–1248 (2005).
9. Nierman, W. C. *et al.* Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*. *Nature* **438,** 1151–1156 (2005).
10. Galagan, J. E. *et al.* Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*. *Nature* **438,** 1105–1115 (2005).
11. Latgé, J. P. The pathobiology of *Aspergillus fumigatus*. *Trends Microbiol.* **9,** 382–389 (2001).
12. Rydholm, C., Szakacs, G. & Lutzoni, F. Low genetic variation and no detectable population structure in *Aspergillus fumigatus* compared to closely related *Neosartorya* species. *Eukaryot. Cell* **5,** 650–657 (2006).
13. Debuchy, R. & Turgeon, B. G. in *The Mycota I: Growth, Differentiation and Sexuality* (eds Kües U. & Fischer, R.) 293–323 (Springer, 2006).
14. Idnurm, A., Walton, F. J., Floyd, A. & Heitman, J. Identification of the *sex* genes in an early diverged fungus. *Nature* **451,** 193–196 (2008).
15. Dyer, P. S., Ingram, D. S. & Johnstone, K. The control of sexual morphogenesis in the Ascomycotina. *Biol. Rev. Camb. Phil. Soc.* **67,** 421–458 (1992).
16. Paoletti, M. *et al.* Mating type and the genetic basis of self-fertility in the model fungus *Aspergillus nidulans*. *Curr. Biol.* **17,** 1384–1389 (2007).
17. Bain, J. M. *et al.* Multilocus sequence typing of the pathogenic fungus *Aspergillus fumigatus*. *J. Clin. Microbiol.* **45,** 1469–1477 (2007).
18. O'Gorman, C. M. & Fuller, H. T. Prevalence of culturable airborne spores of selected allergenic and pathogenic fungi in outdoor air. *Atmos. Environ.* **42,** 4355–4368 (2008).
19. Balajee, S. A., Nickle, D., Varga, J. & Marr, K. A. Molecular studies reveal frequent misidentification of *Aspergillus fumigatus* by morphotyping. *Eukaryot. Cell* **5,** 1705–1712 (2006).
20. Samson, R. A., Hong, S., Peterson, S. W., Frisvad, J. C. & Varga, J. Polyphasic taxonomy of *Aspergillus* section *Fumigati* and its teleomorph *Neosartorya*. *Stud. Mycol.* **59,** 147–203 (2007).
21. Dyer, P. S. in *Sex in Fungi: Molecular Determination and Evolutionary Principles* (eds Heitman, J., Kronstad, J.W., Taylor, J.W. & Casselton, L.A.) 123–142 (ASM Press, 2007).
22. Robert, V. *et al. CBS Yeasts Database* (Centraalbureau voor Schimmelcultures, Utrecht, 2007).
23. Takada, M. & Udagawa, S. A new species of heterothallic *Neosartorya*. *Mycotaxon* **24,** 395–402 (1985).
24. Kwon-Chung, K. J. & Kim, S. J. A second heterothallic *Aspergillus*. *Mycologia* **66,** 628–638 (1974).
25. Pyrzak, W., Miller, K. Y. & Miller, B. L. The mating type protein Mat1-2 from asexual *Aspergillus fumigatus* drives sexual reproduction in fertile *Aspergillus nidulans*. *Eukaryot. Cell* **7,** 1029–1040 (2008).
26. Magee, P. T. & Magee, B. B. Through a glass opaquely: the biological significance of mating in *Candida albicans*. *Curr. Opin. Microbiol.* **7,** 661–665 (2004).
27. Lucas, J. A., Dyer, P. S. & Murray, T. Pathogenicity, host specificity, and population biology of *Tapesia* spp. causal agents of eyespot disease of cereals. *Adv. Bot. Res.* **33,** 225–258 (2000).
28. Ware, S. B. *et al.* Discovery of a functional *Mycosphaerella* teleomorph in the presumed asexual barley pathogen *Septoria passerinii*. *Fungal Genet. Biol.* **44,** 389–397 (2007).
29. Fedorova, N. D. *et al.* Genomic islands in the pathogenic filamentous fungus *Aspergillus fumigatus*. *PLoS Genet.* **4,** doi:10.1371/journal.pgen.1000046 (2008).
30. Taylor, J. W., Geiser, D. M., Burt, A. & Koufopanou, V. The evolutionary biology and population genetics underlying fungal strain typing. *Clin. Microbiol. Rev.* **12,** 126–146 (1999).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank S. A. Balajee for access to unpublished data, C. Duggan for assistance with the Latin diagnosis, and C. O'Connell for taking the scanning electron micrographs. This work was supported by an IRCSET Postgraduate Research Scholarship, an EC Marie Curie Training Fellowship and a grant from the British Mycological Society to C.O'G.

Author Contributions C.O'G., H.T.F. and P.S.D. designed the experiments. C.O'G. performed most of the experiments. C.O'G. and P.S.D. analysed the results and wrote the manuscript. All authors contributed to editing the manuscript.

Author Information DNA sequences have been deposited in GenBank under accession numbers EU541353 and EU541354 (carboxypeptidase-5) and EU541355 (β-tubulin). The holotype of *Neosartorya fumigata* has been deposited in the Herbarium of the Royal Botanic Gardens, Kew, under accession number K(M)159484. The assignment *Neosartorya fumigata* O'Gorman, Fuller & Dyer *sp. nov*. has been deposited in MycoBank under accession number MB 512563. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to C.O'G. (celine.ogorman@ucd.ie) or P.S.D. (paul.dyer@nottingham.ac.uk).

METHODS

Maintenance and crossing of isolates. Single-spore cultures of each isolate were routinely maintained on MEA slopes at 4 °C in the dark. Crosses were set up on Oatmeal agar medium[22], Czapek Dox agar[31] and *Aspergillus* complete medium[8].

Light microscopy and scanning electron microscopy. For micro-morphological examination of *N. fumigata* sexual reproductive structures, mounts were made in clear lactophenol from cultures grown on Oatmeal agar medium[22] and colours were assigned using Methuen colour names and numbers[32]. Dimensions of cleistothecia and ascospores ($n = 20$) were measured with the software package Olympus DP-Soft, version 5.0, linked to an Olympus BX-45 light microscope and an Olympus C-5060 Wide Zoom digital camera. For scoring purposes, cleistothecia were defined as spherical white to yellow structures more than 150 μm in diameter. Scanning electron microscopy was performed with a Jeol JSM-5410LV scanning electron microscope. For preparation for scanning electron microscopy, mature cleistothecia were transferred to aluminium stubs with the use of double-sided adhesive tape. A small drop of 0.05% Tween 80 was added and the cleistothecia were crushed. The suspensions were air-dried and sputter-coated with gold.

DNA extraction. Cultures were grown in 2% liquid malt extract for 5 days at 28 °C, washed with a 0.1 M phosphate buffer (pH 7.0) and ground into a fine powder under liquid nitrogen. Genomic DNA was extracted with a DNeasy Plant Mini Kit (Qiagen) in accordance with the manufacturer's instructions.

Mating-type PCR assay. The mating-type genotype of each isolate was determined with a multiplex-PCR assay as previously described[8], with reaction volumes of 25 μl containing 50–100 ng of DNA. Segregation of mating-type genes in ascospore progeny has previously been used to infer the breeding systems of other filamentous ascomycete species, especially lichen-forming fungi[33,34].

Statistical analyses. The null hypothesis was for a 1:1 ratio of the two mating types, as is characteristic for a randomly mating population of haploid organisms[35]. The hypothesis was tested with χ^2 and contingency χ^2 tests, with $P < 0.05$ considered significant. Fisher's exact test was used instead of the χ^2 test when sample sizes were small, because it is more accurate when expected frequencies are less than five[36].

RAPD-PCR protocol for clone-correction and genetic relatedness. RAPD-PCR was performed with cycling conditions described previously[37]. Four primers were used to fingerprint the 91 *A. fumigatus* isolates: RC08 (5′-GGATGTCGAA-3′)[38], R108 (5′-GTATTGCCCT-3′)[39], R151 (5′-GCTGTAGTGT-3′)[39] and UBC90 (5′-GGGGGTTAGG-3′)[40]. PCR amplifications were performed in 25-μl reaction volumes containing 0.1–1.0 ng of DNA. Each amplification reaction contained 2.5 μl of 10×DyNAzyme buffer, 0.75 U of DyNAzyme II Polymerase (2 U μl^{-1}; Finnzymes OY, Flowgen Laboratories), 0.15 μl of 25 mM dNTP mix (ABgene), 50 μmol of a single ten-base-pair primer and ultra-pure water (Sigma). Amplicons were resolved in 1.5% agarose gels (data not shown). Clear, reproducible RAPD-PCR amplicons from each primer were scored as 0 (amplicon absent) or 1 (amplicon present) in Microsoft Excel. The data from all primers were pooled for each isolate, and a pairwise similarity matrix was calculated with Jaccard's coefficient[41] with the program FREETREE[42] v. 0.9.1.50. A bootstrapped dendrogram with 1,000 replications was produced with NJ Plot. It is noted that the most divergent isolate detected by this analysis, AfIR917, was also the only isolate that failed to amplify a *MAT* product in the multiplex mating-type assay, suggesting sequence divergence in the *MAT* region and that this isolate was not *A. fumigatus sensu stricto*.

RAPD-PCR for genotypic markers. The 12 parental isolates (indicated in Supplementary Table 1) were screened with 37 ten-base RAPD primers (Operon Biotechnologies Kits OP-A, -J, -W and -X) for the amplification of bands suitable for genotypic markers. Five primers (A-05, J-01, W-10, W-19 and X-05) that amplified clear, reproducible bands unique to one of the parental isolates were selected. RAPD-PCR reactions were performed as described above. Segregation of RAPD-PCR markers has previously been used to determine sexual breeding systems in several other ascomycete fungi[33,43–45].

Gene sequencing. The genes encoding β-tubulin and carboxypeptidase-5 were amplified with the primer sets benA (forward, 5′-AATTGGTGCCGCTTTCTGG-3′; reverse, 5′-AGTTGTCGGGACGGAATAG-3′) and cxp (forward, 5′-GAACATTAGCCCCAGTTGAG-3′; reverse primer, 5′-CACTTCTTCTTGCACGTAGTC-3′), respectively (S. A. Balajee, D. Nickle, L. Razai, S. F. Hurst & K. A. Marr, personal communication). The amplified DNA fragments were purified with a QIAquick PCR purification kit (Qiagen). DNA sequencing of the forward strand of each fragment was performed at the Biopolymer Synthesis and Analysis Unit, University of Nottingham. The resulting sequences were aligned in CLUSTALW[46] using the program BioEdit[47].

Measurement of ascospore germination. Using a fine-tipped needle, an intact mature cleistothecium was removed from a 12-month-old culture of an *N. fumigata* AfIR956 × AfRB2 cross. An ascospore suspension was prepared and heat-treated. Ascospore concentration (about 4×10^4 spores ml^{-1}) was determined with a haemocytometer. Defined areas on 2% MEA plates were each spread inoculated with 50 μl of the heat-treated ascospore suspension. Plates were sealed with one layer of Parafilm and incubated at 28 °C in darkness overnight. Coverslips were then placed on the agar surfaces over the defined areas, and levels of ascospore germination were determined. The percentage germination of heat-treated (70 °C, 60 min) ascospores was 50.5 ± 3.4% ($n = 600$; mean ± s.e.m.) after 15 h at 28 °C.

31. Klich, M. A. *Identification of Common* Aspergillus *Species* (Centraalbureau voor Schimmelcultures, Utrecht, 2002).
32. Kornerup, A. & Wanscher, J. H. *Methuen Handbook of Colour* 3rd edn (Eyre Methuen, 1978).
33. Seymour, F. A. *et al.* Breeding systems in the lichen-forming fungal genus *Cladonia*. *Fungal Genet. Biol.* **42,** 554–563 (2005).
34. Honegger, R., Zippler, U., Gansner, H. & Scherrer, S. Mating systems in the genus *Xanthoria* (lichen-forming ascomycetes). *Mycol. Res.* **108,** 480–488 (2004).
35. Milgroom, M. G. Recombination and the multilocus structure of fungal pathogens. *Annu. Rev. Phytopathol.* **34,** 457–477 (1996).
36. Fisher, R. A. *Statistical Methods for Research Workers* 7th edn (Oliver & Boyd, 1938).
37. Murtagh, G. J., Dyer, P. S., McClure, P. C. & Crittenden, P. D. Use of randomly amplified polymorphic DNA markers as a tool to study variation in lichen-forming fungi. *Lichenologist* **31,** 257–267 (1999).
38. Anderson, M. J., Gull, K. & Denning, D. W. Molecular typing by random amplification of polymorphic DNA and M13 southern hybridization of related paired isolates of *Aspergillus fumigatus*. *J. Clin. Microbiol.* **34,** 87–93 (1996).
39. Aufauvre-Brown, A., Cohen, J. & Holden, D. W. Use of random amplified polymorphic DNA markers to distinguish isolates of *Aspergillus fumigatus*. *J. Clin. Microbiol.* **30,** 2991–2993 (1992).
40. Lin, D. *et al.* Comparison of three typing methods for clinical and environmental isolates of *Aspergillus fumigatus*. *J. Clin. Microbiol.* **33,** 1596–1601 (1995).
41. Weising, K., Nybom, H., Wolff, K. & Meyer, W. *DNA Fingerprinting in Plants and Fungi* (CRC Press, 1995).
42. Pavlíček, A., Hrdá, Š. & Flegr, J. FreeTree – a freeware program for construction of phylogenetic trees on the basis of distance data and for bootstrap/jacknife analysis of the tree robustness. Application in the RAPD analysis of genus *Frenkelia*. *Folia Biol. Prague* **45,** 97–99 (1999).
43. Dyer, P. S., Nicholson, P., Rezanoor, H. N., Lucas, J. A. & Peberdy, J. F. Two-allele heterothallism in *Tapesia yallundae*, the teleomorph of the cereal eyespot pathogen *Pseudocercosporella herpotrichoides*. *Physiol. Mol. Plant Pathol.* **43,** 403–414 (1993).
44. Kema, G. H. J., Verstappen, E. C. P., Todorova, M. & Waalwijk, C. Successful crosses and molecular tetrad and progeny analysis demonstrate heterothallism in *Mycosphaerella graminicola*. *Curr. Genet.* **30,** 251–258 (1996).
45. Murtagh, G. J., Dyer, P. S. & Crittenden, P. D. Sex and the single lichen. *Nature* **404,** 564 (2000).
46. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22,** 4673–4680 (1994).
47. Hall, T. A. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp.* **S41,** 95–98 (1999).

LETTERS

Anticipatory haemodynamic signals in sensory cortex not predicted by local neuronal activity

Yevgeniy B. Sirotin[1] & Aniruddha Das[1,2,3,4,5,6]

Haemodynamic signals underlying functional brain imaging (for example, functional magnetic resonance imaging (fMRI)) are assumed to reflect metabolic demand generated by local neuronal activity, with equal increases in haemodynamic signal implying equal increases in the underlying neuronal activity[1–6]. Few studies have compared neuronal and haemodynamic signals in alert animals[7,8] to test for this assumed correspondence. Here we present evidence that brings this assumption into question. Using a dual-wavelength optical imaging technique[9] that independently measures cerebral blood volume and oxygenation, continuously, in alert behaving monkeys, we find two distinct components to the haemodynamic signal in the alert animals' primary visual cortex (V1). One component is reliably predictable from neuronal responses generated by visual input. The other component—of almost comparable strength—is a hitherto unknown signal that entrains to task structure independently of visual input or of standard neural predictors of haemodynamics. This latter component shows predictive timing, with increases of cerebral blood volume in anticipation of trial onsets even in darkness. This trial-locked haemodynamic signal could be due to an accompanying V1 arterial pumping mechanism, closely matched in time, with peaks of arterial dilation entrained to predicted trial onsets. These findings (tested in two animals) challenge the current understanding of the link between brain haemodynamics and local neuronal activity. They also suggest the existence of a novel preparatory mechanism in the brain that brings additional arterial blood to cortex in anticipation of expected tasks.

We have developed a dual-wavelength optical imaging technique to (in effect) image cortical blood volume and oxygenation simultaneously in alert behaving macaques. This technique involves switching rapidly between two wavelengths: 530 nm (green, equally absorbed in oxygenated and deoxygenated haemoglobin, thus measuring total haemoglobin concentration (HbT) or 'blood volume') and 605 nm (red, absorbed about fivefold more strongly in deoxygenated than oxygenated haemoglobin, thus measuring 'oxygenation'[10]

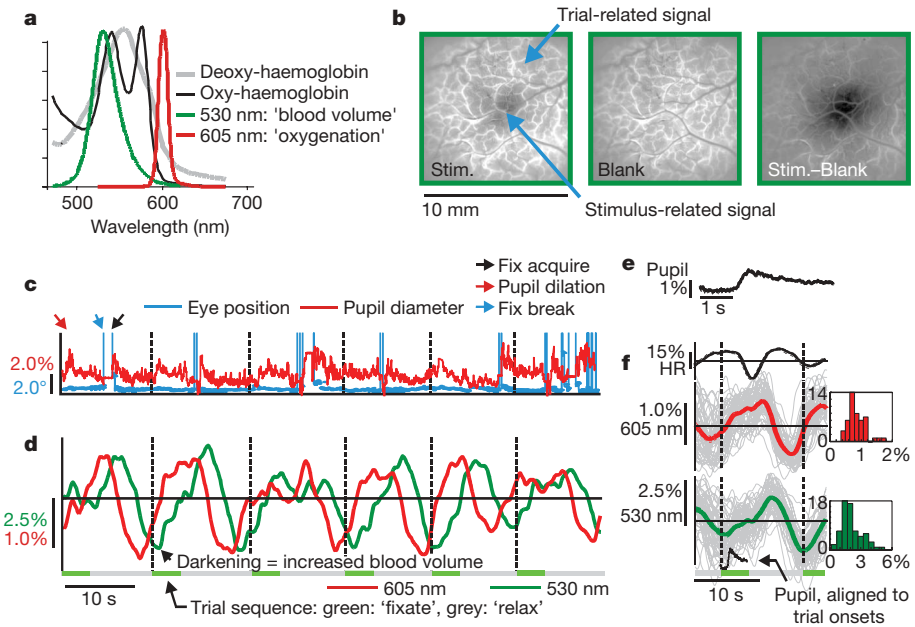


Figure 1 | Periodic fixation tasks evoke stimulus-independent, trial-linked signals even in the dark. **a**, Normalized emission spectra of the two illumination sources (light-emitting diodes), superimposed on *in vitro* absorbance spectra for deoxy- and oxyhaemoglobin[10] (units: $10^4\,cm^{-1}\,M^{-1}$). **b**, Stim.: V1 'blood volume' response to small, brief, visual stimulus presented during periodic fixation trials. 'Blank': signal in trial with no visual stimulus. 'Stim.-Blank': stimulus-specific response. **c**, Eye position and pupil diameter (percentage of mean), consecutive trials. (Vertical dashed lines: trial onsets. Note pupil dilation, fix break, fix acquire, shown for first trial.) Scales are colour-coded. **d**, Cortical signals, colour-coded by imaging wavelength. **e**, **f**, Trial-triggered averages (grey lines, individual trials; thick lines, mean ± s.e.m., $n = 51$: at 605 nm, mean peak-to-peak amplitude = 1.19% ± 0.08; at 530 nm, 3.47% ± 0.21). Population histograms: at 605 nm, mean = 0.86%, s.d. = 0.29, $N = 47$ experiments; at 530 nm, mean = 2.17%, s.d. = 0.97, $N = 66$.

[1]Department of Neuroscience, [2]Department of Psychiatry, [3]W. M. Keck Center on Brain Plasticity and Cognition, [4]Mahoney Center for Brain and Behavior, [5]Department of Biomedical Engineering, Columbia University, New York, New York 10027, USA. [6]New York State Psychiatric Institute, 1051 Riverside Drive, Unit 87, New York, New York 10032, USA.

(Fig. 1a and Supplementary Methods)). While imaging V1 in animals performing periodic visual tasks, we observed a hitherto unknown stimulus-independent haemodynamic signal that appeared to entrain to trial timing (Fig. 1b).

To study this trial-related signal in isolation, we developed a task that minimized visual input while preserving trial timing. In an otherwise completely dark room, the animal was required to fix its gaze periodically on a tiny fixation point for juice reward (point size about 1–2 arcmin, that is, about 1–2 cone diameters). The fixation point stayed on continuously, switching between two equiluminant colours to cue the animal to 'fixate' or 'relax'. It was akin to seeing nothing apart from one single twinkling star in an otherwise black night sky. Our two rhesus macaque monkeys ('V' and 'S') learned the task correctly, as evidenced by their fixation patterns (Fig. 1c). Both monkeys performed long sequences of correct trials, consistently holding fixation during 'fixate' periods and taking fixation breaks, if any, only during 'relax' periods.

On imaging V1 while the animals performed this task, we observed robust haemodynamic signals at the trial frequency, even though the animals were in virtually total darkness and foveal V1, the only region receiving visual input from the fixation point, lay outside our imaging area. These periodic fluctuations were seen in both the 'blood volume' (530 nm) and 'oxygenation' signals (605 nm; Fig. 1d, f). They were accompanied by periodic changes in heart rate[11] and systematic pupil dilation[12] on trial onset, suggesting a rhythmic state of alertness synchronized to each trial (Fig. 1c–f).

We wanted to determine the relation between these trial-linked haemodynamic signals and V1 neuronal activity. A crucial assumption in most brain imaging studies is that haemodynamic signals are caused by local neuronal responses through a uniform underlying mechanism[1–6] (but see refs 13, 14). In particular, brain images are routinely used to infer changes in local neuronal activity by fitting the imaging signal with some standard causal haemodynamic kernel. To reveal neuronal mechanisms underlying V1 haemodynamics, we obtained both trial-related and visually evoked optical imaging signals concurrently with electrode recordings across V1 (Supplementary Fig. 1 and Supplementary Table 1). At each site, in alternating blocks (20–40 trials each) while the animal performed the same fixation task, we presented either vigorous visual stimuli or no stimuli at all. For each data set we then used an optimization routine to calculate the causal kernel that 'best' fitted haemodynamics to concurrent neuronal signals (Supplementary Fig. 2), and tested whether this 'best' kernel could reliably predict haemodynamics.

To get measures of neuronal activity for this analysis, we separated the electrode recordings into multi-unit spiking (MUA) and local field potential (LFP) (Supplementary Fig. 1 and Supplementary Methods). As expected, visual stimulation evoked vigorous responses in both MUA and LFP (Fig. 2a). The stimulus-evoked LFP responses could

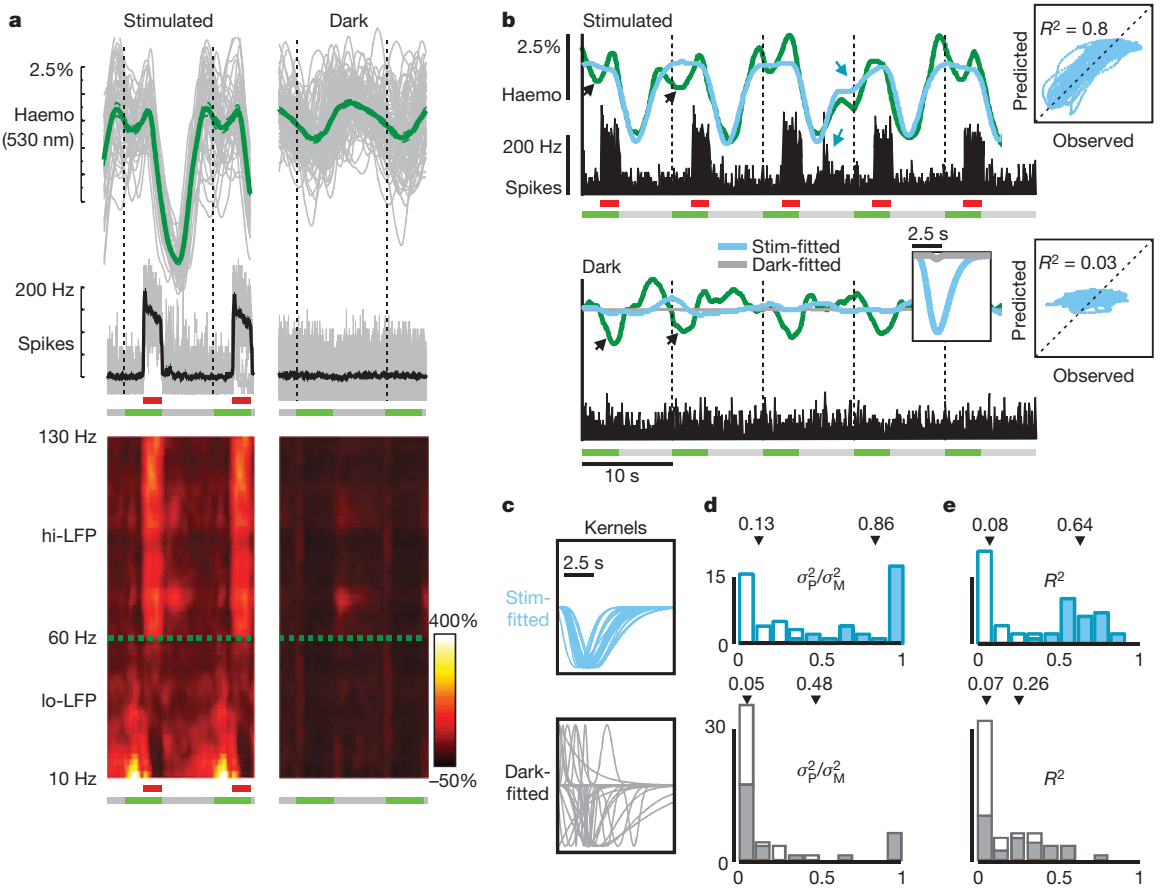


Figure 2 | Local neuronal activity predicts visually driven, but not trial-related, haemodynamics. **a**, Trial-triggered mean haemodynamic ('blood volume') and electrophysiological signals comparing stimulus-driven and dark-room responses, representative experiment. LFP power spectrum (bottom) normalized to pre-stimulus dark power (2-Hz resolution). **b**, Comparing measured haemodynamics (green) with optimal predictions from concurrent spiking, same experiment. Blue, grey: using kernels (inset) obtained by fitting stimulated or dark-room signals respectively. (The same colour code is used throughout. Dark-room kernel and prediction almost indistinguishable from a flat line; prediction shown for dark only, to avoid clutter.) Black arrows: trial-related activity not predicted in either the stimulated or dark-room trials. Blue arrows: random bursts of neuronal activity generate matching deflections in the predicted and observed trace. Right: scatter-plots and R^2 values of observed versus predicted haemodynamics using stimulus-based predictors. **c**, Optimal kernels across days (amplitude normalized for comparison; $N = 28$ recording sites). **d**, **e**, Descriptive statistics of spike-based fits. Top: stimulus-based prediction; bottom: dark-room based. **d**, Ratio of variance between predicted and measured signals (σ_P^2/σ_M^2). **e**, R^2 statistic: open versus closed bars represent dark-room versus stimulus-driven sessions, respectively. Arrows mark population means.

be empirically separated into two distinct frequency bands (Fig. 2a, bottom). The high-frequency band (hi-LFP: 66–130 Hz, avoiding 60 Hz), like MUA, showed crisp, visually evoked responses. The low-frequency band (lo-LFP: 10–56 Hz), also showed a robust signal but with no apparent correlation with visual stimulation. Our empirically defined LFP bands match categories defined through previous work. The hi-LFP matches a frequency band ('high gamma') that is shown to correlate well with stimulus-evoked spiking and haemodynamics[2,15,16]. The lo-LFP—often separated into finer frequency bands[15,16]—is believed to have a very different relationship with other brain signals[15,16]. We therefore decided to test the three neuronal signal types, MUA, hi-LFP and lo-LFP, independently for their ability to predict concurrently recorded haemodynamics reliably. These tests were conducted separately for 'blood volume' and 'oxygenation' signals.

Visually driven MUA and hi-LFP predicted the simultaneously recorded haemodynamic signals very well both in amplitude and time course (Fig. 2b–e and Supplementary Fig. 3b–e). Further, the optimal kernels obtained by fitting these signals were consistent in shape across all recording sites (Fig. 2c and Supplementary Fig. 3c, top); kernels from any given experiment predicted visually evoked responses in all other experiments with almost equal accuracy, attesting to their remarkable reliability (Supplementary Fig. 4). In sharp contrast, the same kernels, when convolved with dark-room MUA or hi-LFP, were uniformly poor at predicting trial-related haemodynamics, in both amplitude and temporal correlation (Fig. 2b–e and Supplementary Fig. 3b–e). The latter finding ($R^2 \sim 0.08$, MUA; 0.06, hi-LFP) specifically implies that there is no consistent temporal relation between predicted and measured haemodynamics, independent of amplitude. This poor predictability was particularly striking because the trial-related haemodynamic signal amplitudes were almost comparable to those of responses to vigorous visual stimulation (37% at 'blood volume', 530 nm; 57% at 'oxygenation', 605 nm (Supplementary Fig. 5)). To check whether trial-related haemodynamics could still be predicted reliably by concurrent neuronal recordings but through kernels of a different shape, we fitted dark-room MUA and hi-LFP to dark-room haemodynamics. These 'best' dark-room kernels were highly variable among recording sites and, again, consistently failed to predict trial-related haemodynamics (Fig. 2c–e and Supplementary Fig. 3c–e, bottom). The same overall pattern of results was seen for both 'blood volume' and 'oxygenation' signals (Supplementary Fig. 6).

These results provide compelling evidence that visually evoked haemodynamic signals are very well predicted by established measures of local neuronal activity (MUA, hi-LFP) through a causal kernel that is uniform across experiments. Such a model fails profoundly, however, to predict the trial-related signals. Therefore any neuronal mechanisms underlying trial-related haemodynamics appear to be distinct from those typically assumed to underlie neurovascular coupling.

Unlike MUA or hi-LFP, lo-LFP—whether treated as a whole or separated into finer frequency bands—failed to show any consistent relationship with haemodynamics. These signals gave highly variable 'optimal kernels' when fitted with concurrent haemodynamics either under visual driving or in the dark, with uniformly poor predictions of haemodynamics (Supplementary Figs 7 and 13).

Next, we characterized the novel trial-related haemodynamic signal in terms of its temporal relation to trial timing. To determine whether our observed signals are linked specifically with trial timing and not a result of some unrelated intrinsic oscillatory process[17], we examined how the signals adapted to different trial periods. Our results provided compelling evidence that the signals are linked predictively to trial onsets. This was seen both in the signal shapes at each trial period and their anticipatory timing on switching trial period.

We found that the trial-related signals stretched elastically to match each tested trial period (Fig. 3 and Supplementary Fig. 8a; tested 6-s to 30-s trial periods). In particular, the shape of the 'blood volume' signal

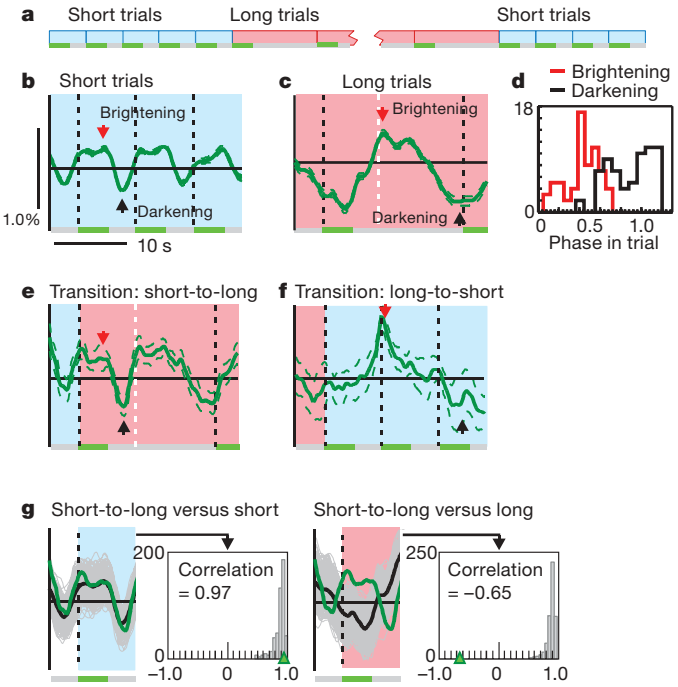


Figure 3 | Trial-related haemodynamic signals entrain to anticipated trial onsets, stretching to conform to the trial period. **a**, Dark-room fixation trials, each with the same 4-s 'fixate' epochs but in blocks of different trial periods, short (8 s, blue) and long (20 s; pink) (same colour scheme in other panels and Fig. 4). **b**, **c**, 'Blood volume' signal, short and long trials respectively. Averages triggered on trial onsets, ± s.e.m. (**b**: $N = 220$ trials; **c**: $N = 179$). Arrowheads: peak brightening (red), darkening (black). Dotted white line in **c** shows where short-trial onset would have occurred. **d**, Population histograms of peak brightening and darkening in units of trial phase (0, trial start; 1, trial end, start of next trial; brightening, mean = 0.43, s.d. = 0.16; darkening, mean = 0.89, sd = 0.21; $N = 66$; trial periods ranged from 6 to 30 s). **e**, **f**, Signal at transitions between trial periods: short-to-long (**e**, $N = 16$ trials, mean ± s.e.m.), and long-to-short (**f**, $N = 10$); arrowheads are aligned, in each case, to the panel above for comparison of signal features. Dotted white line as in **c**. **g**, Left: short-to-long transition trial (green) is statistically indistinguishable from other short trials over one short-trial period (blue background). (Bootstrap analysis. Green: mean transition trial; grey: means of 500 random $N = 16$-trial subsets of the short trials to match statistics of transition trial; black: grand mean of all short trials, same as **b**; inset histogram: correlation of random subsets with grand mean; arrowhead: correlation of transition trial with grand mean = 0.97.) Right: short-to-long transition response is distinct from random $N = 16$-trial subsets of long trials. Same conventions as on the left, with the correlation coefficients being calculated, again, over the duration of one short period (pink background).

always stretched so as to start darkening (increasing haemoglobin) during the 'relax' period—before the onset of the next trial—reaching a peak darkening close to the onset of the next 'fixate' period (Figs 1 and 3a–d). This elastic pattern of trial-locked haemodynamics—in which signals begin changing before trial onsets—cannot be explained by mechanisms that involve a causal kernel triggered on trial start. This can be demonstrated by comparison with responses to (brief, intense) visual stimulation of the same duration as the 'fixate' period, where the stereotyped response shape, with abrupt onset and fixed width after stimulus presentation, is independent of trial period (Supplementary Fig. 8b; quantitative model, Supplementary Fig. 8c). The trial-related signal is thus unlikely to be due to neuronal signals active only during the cued 'fixate' period (for example, the presumed time course of 'attention'[18]).

On switching trial timing unexpectedly after the monkey had established a rhythm of 10–20 correct trials at a given period, haemodynamic signals continued to oscillate at the earlier period for a couple of trials before entraining to the new one (Fig. 3e, f). This

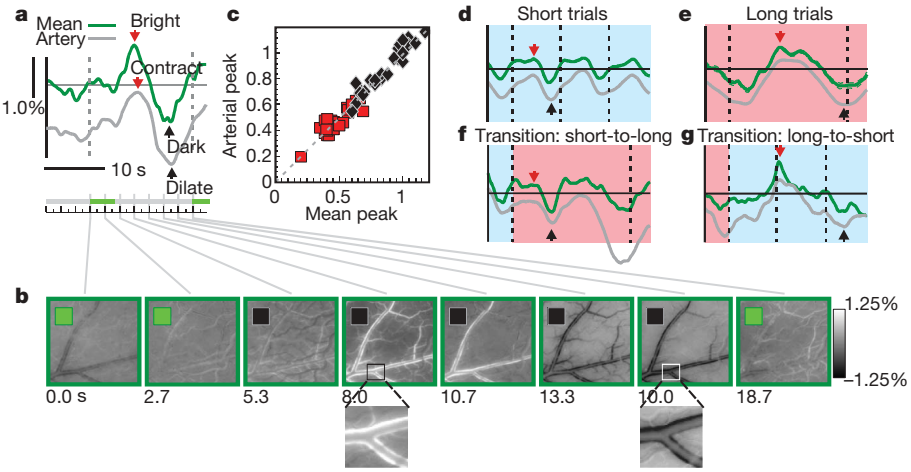


Figure 4 | Mean 'blood volume' signal is closely matched, temporally, by V1 arterial contraction–dilation cycle. **a**, Mean trial-triggered signals and **b** individual frames showing fractional signal change relative to trial-mean image. Inset square: green, 'fixate'; black, 'relax'. Magnified sections show arterial contraction (white walls) and dilation (black walls). Grey trace in **a**: arterial signal relative to 'parenchyma baseline' (Supplementary Fig. 11, method for calculating arterial signal; arterial trace shifted vertically from overall mean for visibility). **c**, Timing of peak arterial contraction (dilation), as phase within trial, matches peak brightening (darkening) of mean signal: red square (black diamond) respectively. **d–g**, Arterial signal (grey) closely matches mean signal (green) for different trial periods (**d**, **e**) and at transitions between periods (**f**, **g**); same experiment, conventions as in Fig 3b, c, f, g (traces shifted vertically for visibility). Note close matches between corresponding peaks and troughs (arrows), indicated as in Fig. 3.

occurred even though the animal itself picked up the new trial pace immediately, holding and breaking fixation at the new rhythm right after the switch (that is, clearly having noticed the new pace of fixation cues). Thus, on switching from short to long trials the measured signals showed a peak darkening at the short-trial spacing even though the animal was fixating correctly at the longer period (Fig. 3e). Similarly, on switching from long to short trials the cortical signal continued at its previous slower pace for one long period, overriding the first few short trials (Fig. 3f). The response shape observed on transition trials closely resembled pre-transition responses for the duration of the pre-transition trial period, while being very poorly matched to the post-transition trial shape, suggesting that the underlying neuronal mechanism continued to 'anticipate' the pre-transition trial timing (Fig. 3g and Supplementary Fig. 9). Further, the trial-related signal timing was correlated specifically with trial onsets and not with reward[19]: the peak darkening position remained unaffected on delaying the reward associated with each trial (Supplementary Fig. 8d).

Finally, images of the cortical surface suggested that the trial-related signals involve the local vasculature rather than being a systemic trial-locked autonomic (for example, cardiac) response[20]. These images revealed a dramatic contraction–dilation cycle in V1 arteries, evidenced by a prominent brightening, followed by darkening of the arterial walls relative to the 'parenchyma' baseline (Fig. 4a, b and Supplementary Fig. 10; Supplementary Fig. 11 indicates how arteries, veins and 'parenchyma' are distinguished and how the arterial signal is measured). This arterial signal had a timing that closely matched the overall timing of the mean 'blood volume', with peaks of arterial contraction and dilation coinciding with peaks of brightening (decreased haemoglobin) and darkening (increased haemoglobin), respectively (Fig. 4a, c–g). The arterial cycle stretched elastically to fit trial periods, matching the shape of the mean signal (Fig. 4d, e). Further, on switching trial periods, the arterial cycle showed an anticipatory dilation that was well synchronized with the anticipatory increase in 'blood volume' seen in the mean signal (Fig. 4f, g). This local arterial cycle may thus be the specific mechanism generating trial-related increases in V1 'blood volume' in anticipation of visual tasks. Further, the arterial cycle is seen in V1 only for visual tasks and is likely not a passive consequence of trial-locked changes in heart rate or blood pressure[20]. We found no V1 arterial pumping or trial-related changes in V1 'blood volume' in a periodic auditory control task, despite the presence of periodic changes in heart rate and pupil dilation very similar to those seen in our visual task (Supplementary Fig. 12).

Our findings have two major implications: one for the interpretation of brain imaging[21], the other advancing our knowledge of brain mechanisms underlying anticipation. First, the interpretation of fMRI[22], for example, through general linear modelling[23], typically makes the crucial assumption of a uniform linear predictive relationship between neuronal and haemodynamic signals. We show that this model is valid for visually evoked signals, but that it fails profoundly to predict another class of signals, of almost comparable magnitude and behaviourally linked structure. These results raise the further possibility that there may be other, hitherto uncovered exceptions[13,14] to the assumption that haemodynamic signals uniformly imply equivalent underlying neuronal activity. Second, the predictive timing and arterial contraction–dilation cycle that we observe in the trial-related haemodynamic signal suggests that it could reflect a novel anticipatory brain mechanism. This mechanism could play the role of preparing cortex for anticipated tasks by bringing additional arterial blood in time for task onsets. The question of the mechanism driving this signal (for example, distal neuromodulatory control of cerebral arteries?) and its functional consequences remains a challenge for future investigations.

METHODS SUMMARY

Results were obtained using continuous, dual-wavelength intrinsic-signal optical imaging and electrode recording in two monkeys engaged in either visual fixation tasks or auditory control tasks. Standard alert-monkey optical imaging techniques[24] were used to record the intrinsic cortical signal, continuously, through a clear silicone artificial dura and glass-fronted recording chamber implanted over the V1 of the animals. The primary innovation here consisted of our using two imaging wavelengths. Two arrays of fast, high-intensity light-emitting diodes at the two wavelengths (530 nm, 605 nm) were switched on and off alternately in synchrony with the camera, thus illuminating the brain surface alternately with each wavelength on successive camera frames (15 frames per second). The illumination alternated much faster than typical haemodynamic signal timescales giving, in effect, simultaneous optical imaging at both wavelengths at 7.5 frames per second. Increased absorption (darkening) at 530 nm indicated an increase in total haemoglobin ('blood volume'). Increased absorption at 605 nm primarily indicated an increase in deoxyhaemoglobin, from a combination of increased deoxygenation and blood volume.

For the dark-room fixation task, in a completely dark room, with a mask covering even the stimulus presentation monitor, the animal was cued to fixate or relax by the colour of a fixation point visible through a pinhole in the mask (size 1–2 arcmin), typically switching between equiluminant green ('fixate') and red ('relax'). We dark-adapted alongside the animal to confirm that nothing else

was visible. Control experiments confirmed that the trial-related signal was independent of the brightness (range: 10×), colour and size (range: 25× in area) of the fixation point.

All experimental procedures were performed in accordance with the National Institutes of Health Guide for the Care and Use of Laboratory Animals and were approved by the Institutional Animal Care and Use Committees of Columbia University and the New York State Psychiatric Institute.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 29 May; accepted 24 November 2008.

1. Logothetis, N. K. & Wandell, B. A. Interpreting the BOLD signal. *Annu. Rev. Physiol.* **66,** 735–769 (2004).
2. Logothetis, N. K., Pauls, J., Augath, M., Trinath, T. & Oeltermann, A. Neurophysiological investigation of the basis of the fMRI signal. *Nature* **412,** 150–157 (2001).
3. Vanzetta, I. & Grinvald, A. Increased cortical oxidative metabolism due to sensory stimulation: implications for functional brain imaging. *Science* **286,** 1555–1558 (1999).
4. Ugurbil, K., Toth, L. J. & Kim, D.-S. How accurate is magnetic resonance imaging of brain function. *Trends Neurosci.* **26,** 108–114 (2003).
5. Shulman, R. G., Rothman, D. L., Behar, K. L. & Hyder, F. Energetic basis for brain activity: implications for neuroimaging. *Trends Neurosci.* **27,** 489–495 (2004).
6. Heeger, D. J., Huk, A. C., Geisler, W. S. & Albrecht, D. G. Spikes versus BOLD: what does neuroimaging tell us about neuronal activity? *Nature Neurosci.* **3,** 631–633 (2000).
7. Berwick, J. *et al.* Hemodynamic response in the unanesthetized rat: intrinsic optical imaging and spectroscopy of the barrel cortex. *J. Cereb. Blood Flow Metab.* **22,** 670–679 (2002).
8. Goense, J. B. M. & Logothetis, N. K. Neurophysiology of the BOLD fMRI signal in awake monkeys. *Curr. Biol.* **18,** 631–640 (2008).
9. Hillman, E. M. C. Optical brain imaging in vivo: techniques and applications from animal to man. *J. Biomed. Opt.* **12,** 051402–1–051402-28 (2007).
10. Prahl, S. Tabulated molar extinction coefficient for hemoglobin in water. <http://omlc.ogi.edu/spectra/hemoglobin/summary.html> (2007).
11. van der Molen, M. W., Boomsma, D. I., Jennings, J. R. & Nieuwboer, R. T. Does the heart know what the eye sees? A cardiac/pupillometric analysis of motor preparation and response execution. *Psychophysiology* **26,** 70–80 (1989).
12. Beatty, J. Task-evoked pupillary responses, processing load and the structure of processing resources. *Psychol. Bull.* **91,** 276–292 (1982).
13. Nir, Y. *et al.* Coupling between neuronal firing rate, gamma LFP, and BOLD fMRI is related to interneuronal correlations. *Curr. Biol.* **17,** 1275–1285 (2007).
14. Maier, A. *et al.* Divergence of fMRI and neural signals in V1 during perceptual suppression in the awake monkey. *Nature Neurosci.* **11,** 1193–1200 (2008).
15. Niessing, J. *et al.* Hemodynamic signals correlate tightly with synchronized gamma oscillations. *Science* **309,** 948–951 (2005).
16. Belitski, A. *et al.* Low-frequency local field potentials and spikes in primary visual cortex convey independent visual information. *J. Neurosci.* **28,** 5696–5709 (2008).
17. Mayhew, J. E. W. *et al.* Cerebral vasomotion: a 0.1-Hz oscillation in reflected light imaging of neural activity. *Neuroimage* **4,** 183–193 (1996).
18. Ress, D., Backus, B. T. & Heeger, D. J. Activity in primary visual cortex predicts performance in a visual detection task. *Nature Neurosci.* **3,** 940–945 (2000).
19. Shuler, M. G. & Bear, M. F. Reward timing in the primary visual cortex. *Science* **311,** 1606–1609 (2006).
20. Franceschini, M. A., Joseph, D. K., Huppert, T. J., Diamond, S. G. & Boas, D. A. Diffuse optical imaging of the whole head. *J. Biomed. Opt.* **11,** 054007 (2006).
21. Attwell, D. & Iadecola, C. The neural basis of functional brain imaging signals. *Trends Neurosci.* **25,** 621–625 (2002).
22. Ogawa, S. *et al.* Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *Proc. Natl Acad. Sci. USA* **89,** 5951–5955 (1992).
23. Glover, G. H. Deconvolution of impulse response in event-related BOLD fMRI. *Neuroimage* **9,** 416–429 (1999).
24. Shtoyerman, E., Arieli, A., Slovin, H., Vanzetta, I. & Grinvald, A. Long-term optical imaging and spectroscopy reveal mechanisms underlying the intrinsic signal and stability of cortical maps in V1 of behaving monkeys. *J. Neurosci.* **20,** 8111–8121 (2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank: K. Korinek for designing and fabricating much of the dual-wavelength optical imaging hardware; P. P. Mitra for the suggestion of making continuous recordings and the use of the Chronux analysis software; E. M. C. Hillman for insights into brain haemodynamic mechanisms; C. Ma, G. Cantone, J. Ordinario, E. Glushenkova, W. Zhang, and M. Bucklin for help with recordings; E. Seidemann and R. Siegel for technical help during our initial setup; C. D. Gilbert and members of the Mahoney Center and the Center for Theoretical Neuroscience at Columbia University for comments on the manuscript. The work was supported by the Keck foundation, grants from the National Institutes of Health, the Klingenstein Foundation, the Gatsby Initiative in Brain Circuitry and the Dana Foundation to A.D. and a National Research Service Award to Y.B.S.

Author Contributions The two co-authors collaborated on almost every aspect of this work.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to A.D. (ad2069@columbia.edu).

METHODS

Tasks: visual fixation. Two monkeys were trained on a variety of visual tasks with a common periodic fixation schedule cued by fixation point colour. The tasks required only passive fixation during 'fixation on' for juice reward (fixation window: 0.5° radius; monitor distance: 133 cm; fix duration: 4 s within trials of duration ranging from 6 to 30 s; trial duration typically held fixed for a given experiment; on some experiments, trial timing switched in blocks between two or three specific values; in other, control experiments, it was randomized by drawing numbers from a homogenous set). Eye fixation and pupil diameter recorded using an infrared eye tracker[25].

Tasks: auditory control. See Supplementary Fig. 10, auditory pitch discrimination task, in a completely dark room (lacking even the fixation point). The trial sequence was as follows: animal pulls lever (start trial) → fixed delay (range 4–10 s in different experiments) → auditory tone onset → delay (typically 4 s) → tone changes pitch, cue to release lever as quickly as possible (200 ms), for juice reward. Once the animals learned this task they performed trials in rapid succession. Thus we could determine task periodicity by setting the initial delay from lever pull to tone on. The monkey typically continued looking in the general direction of the fixation point (even though none was present), allowing us to track pupil dilation with the infrared camera.

Optical imaging: surgery, recording chambers, artificial dura. After the monkeys were trained on visual fixation tasks, craniotomies were performed over the animals' V1, and glass-windowed stainless steel recording chambers were implanted, under surgical anaesthesia, using standard sterile procedures[24,26], so as to image an area of V1 of about 10 mm covering a visual eccentricity of about 1–5°. The exposed dura was resected and replaced with a soft, clear silicone artificial dura. After the animals had recovered from the surgery, cortical activity from their V1 was optically imaged, routinely, while they engaged in relevant behavioural tasks. Recording chambers and artificial dura were fabricated in our laboratory using published methods[27].

Optical imaging: hardware. The camera was a Dalsa 1M30P (binned to 256 pixels × 256 pixels, 15 frames per second). The frame grabber was an Optical PCI Bus Digital (Coreco Imaging). The software was developed in our laboratory based on a system by V. Kalatsky[28]. The illumination was high-intensity light-emitting diodes (Agilent Technologies, Purdy Technologies), with emission wavelengths centred at 530 and 605 nm, filtered through small individual interference filters (Omega Optical). The lens was a 'macroscope' of back-to-back camera lenses[29] focused on the cortical surface. Image acquisition was continuous, simultaneously recording signals from camera, trial timing and behavioural data (trial onset, stimulus onset, identity and duration, etc., eye position, pupil size, timing of fixation breaks, fixation acquisitions, trial outcomes). Data were analysed offline using custom software (MATLAB).

Optical imaging: image processing. All images were first corrected for residual brain movements by aligning each frame to the first frame (shift + rotation[30]), using blood vessels. Signal means (for example, in Fig. 1b) were obtained by averaging signals over the full area, then dividing by the trial-mean of this average to give the percentage signal change as a function of time in a trial. Images of cortical signal (Figs 1a and 4b and Supplementary Movie) were obtained by aligning image sequences to a selected time point (for example, trial onset) and averaging, frame by frame, across the set of all correct trials. This gave 'movies' of cortical activity at the camera frame rate (7.5 frames per second at each wavelength). For images of stimulus-evoked responses (Fig. 1a) (stimulus: 0.25° bar, flashed on for 1 s at the start of each fixation trial), each frame in the movie was then divided, pixel by pixel, by the mean pre-stimulus image (five frames preceding time = 0 ms). The 'Blank' response in Fig. 1b was obtained the same way, at the same time point (3.3 s after stimulus onset), but on a fixation trial with no stimulus. For imaging the trial-related signal (Fig. 4b and Supplementary Movie), each frame in the movie was divided, pixel by pixel, by the trial mean (average of all images over one trial duration), rather than pre-stimulus image, to give the image of fractional signal change relative to the trial mean (Fig. 4b and Supplementary 9). To get time courses of blood-vessel signal relative to the mean (Fig. 4c), the signal was measured along test lines, sampling veins, arteries and 'parenchyma', and the 'parenchyma baseline' regressed away (Supplementary Fig. 9). To get movies of stimulus-evoked activity, similar movies were obtained of cortical activity aligned to stimulus onset, both for trials with stimulus present ('stimulated') and absent ('blank'). Movie frames were divided by the pre-stimulus baseline of three to five frames immediately preceding stimulus onset to get the overall change in cortical activity after stimulus (Supplementary Fig. 1b). We only included trials where the animal maintained fixation correctly.

Visual stimuli for comparing stimulated versus dark-room responses. Gratings were optimized to stimulate the recorded location, typically 100% contrast, 4c/deg, drifting at 4° per second.

Electrophysiology: hardware, electronics and analysis. Extracellular electrode recordings (plastic-coated tungsten: FHC or tungsten in glass; impedances 300–800 kΩ; Plexon amplifier and recording software) were conducted simultaneously with optical imaging (Supplementary Fig. 1). Penetrations were distributed over imaged V1. Recording sites sampled cortical depths starting from most superficial to about 1500 μm below the pial surface at steps of 200–400 μm (Supplementary Table 1). The electrode signal was split into 'spiking' (100 Hz to 8 KHz bandpass) and LFP (0.7–170 Hz). MUA events were defined as each negative-going crossing of a threshold equal to about four times the root mean square of the baseline obtained while the animal looked at a grey screen (Supplementary Fig. 1).

Electrophysiology: data processing. MUA were binned into 16.67-ms bins and aligned to the haemodynamic traces using simultaneously recorded synch events. LFP data were spectrally decomposed using mtspecgramc (Chronux Toolbox for MATLAB; sliding window of 1 s, a step size of 250 ms, frequency range 10–130 Hz) and interpolated into a continuous power spectrum aligned to the haemodynamic traces. Two-dimensional spectrograms (Fig. 2a bottom) show the trial-triggered mean LFP power normalized by the mean pre-trial power in the dark signal (2 Hz frequency resolution). The LFP time course (Supplementary Figs 4 and 6) shows the bandpass-filtered power, integrated over each relevant frequency band ('low-frequency' 10–56 Hz, or 'high-frequency' 66–130 Hz, avoiding 60 Hz).

Electrophysiology: fitting to haemodynamics. For each electrophysiological measure (MUA, low-frequency LFP, high-frequency LFP) the 'best' kernel predicting haemodynamics from concurrent electrophysiology was calculated (Supplementary Fig. 3). Correct trials were extracted from the continuous time series and concatenated into a continuous series. We modelled the haemodynamic response function (HRF) as a gamma kernel: $HRF(t, T, W, A) = A \times (t/T)^{\alpha} \times \exp[(t-T)/-\beta]$, where $\alpha = (T/W)^2 \times 8.0 \times \log(2.0)$, $\beta = W^2/T/8.0/\log(2.0)$, A is the amplitude, T is the time to peak and W is the full width at 75% maximum. We fit the kernel parameters using a downhill simplex algorithm (fminsearch, MATLAB) by comparing the actual haemodynamic response obtained during stimulated trials to that predicted from a convolution of the HRF with the corresponding spike or gamma-band (66–130 Hz) LFP trace. The algorithm reliably converged to similar temporal HRF parameters across all days ($T = 2.50$ (0.08) s, $W = 1.68$ (0.06) s). The proportion of the variance in the haemodynamic responses explained by neuronal activity was quantified using the R^2 statistic from linear regression of the predicted haemodynamic trace to the observed trace both for the stimulated and the dark-room trials.

Controls for trial timing. We performed control experiments to verify that the observed signals were tied specifically to task-related trial onsets, independent of other timing signals. We confirmed that the signal periodicity was not linked to the animal either acquiring or breaking fixation: the two time points, during each trial, with any change in light on the retina (albeit minuscule). This also ruled out any links to extra-retinal fixation-related V1 activity. We controlled for the rhythmic pupil dilations, that is, for the possibility that cortical signals were being evoked by the accompanying pulse of extra light. Giving the animal simulated pupil dilations—a bright flash in the fixation point—evoked no cortical response. As a control for the possibility that the signal was an accidental match to ongoing oscillations, we introduced 20% jitter in trial timing; the signal specifically entrained to trial onsets.

25. Matsuda, K., Nagami, T., Kawano, K. & Yamane, S. A new system for measuring eye position on a personal computer. *Soc. Neurosci. Abstr.* **26,** 744.2 (2000).
26. Das, A. & Gilbert, C. D. Long-range horizontal connections and their role in cortical reorganization revealed by optical recording of cat primary visual cortex. *Nature* **375,** 780–784 (1995).
27. Arieli, A., Grinvald, A. & Slovin, H. Dural substitute for long-term imaging of cortical activity in behaving monkeys and its clinical implications. *J. Neurosci. Methods* **114,** 119–133 (2002).
28. Kalatsky, V. A. & Stryker, M. P. New paradigm for optical imaging: temporally encoded maps of intrinsic signals. *Neuron* **38,** 529–545 (2003).
29. Ratzlaff, E. H. & Grinvald, A. A tandem-lens epifluorescence macroscope: hundred-fold brightness advantage for wide-field imaging. *J. Neurosci. Methods* **36,** 127–137 (1992).
30. Lucas, B. D. & Kanade, T. in *Proc. 7th Int. Joint Conf. Artificial Intelligence* 674–679 (1981).

LETTERS

A core gut microbiome in obese and lean twins

Peter J. Turnbaugh[1], Micah Hamady[3], Tanya Yatsunenko[1], Brandi L. Cantarel[5], Alexis Duncan[2], Ruth E. Ley[1], Mitchell L. Sogin[6], William J. Jones[7], Bruce A. Roe[8], Jason P. Affourtit[9], Michael Egholm[9], Bernard Henrissat[5], Andrew C. Heath[2], Rob Knight[4] & Jeffrey I. Gordon[1]

The human distal gut harbours a vast ensemble of microbes (the microbiota) that provide important metabolic capabilities, including the ability to extract energy from otherwise indigestible dietary polysaccharides[1–6]. Studies of a few unrelated, healthy adults have revealed substantial diversity in their gut communities, as measured by sequencing 16S rRNA genes[6–8], yet how this diversity relates to function and to the rest of the genes in the collective genomes of the microbiota (the gut microbiome) remains obscure. Studies of lean and obese mice suggest that the gut microbiota affects energy balance by influencing the efficiency of calorie harvest from the diet, and how this harvested energy is used and stored[3–5]. Here we characterize the faecal microbial communities of adult female monozygotic and dizygotic twin pairs concordant for leanness or obesity, and their mothers, to address how host genotype, environmental exposure and host adiposity influence the gut microbiome. Analysis of 154 individuals yielded 9,920 near full-length and 1,937,461 partial bacterial 16S rRNA sequences, plus 2.14 gigabases from their microbiomes. The results reveal that the human gut microbiome is shared among family members, but that each person's gut microbial community varies in the specific bacterial lineages present, with a comparable degree of co-variation between adult monozygotic and dizygotic twin pairs. However, there was a wide array of shared microbial genes among sampled individuals, comprising an extensive, identifiable 'core microbiome' at the gene, rather than at the organismal lineage, level. Obesity is associated with phylum-level changes in the microbiota, reduced bacterial diversity and altered representation of bacterial genes and metabolic pathways. These results demonstrate that a diversity of organismal assemblages can nonetheless yield a core microbiome at a functional level, and that deviations from this core are associated with different physiological states (obese compared with lean).

We characterized gut microbial communities in 31 monozygotic twin pairs, 23 dizygotic twin pairs and, where available, their mothers ($n = 46$) (Supplementary Tables 1–5). Monozygotic and dizygotic co-twins and parent–offspring pairs provided an attractive model for assessing the impact of genotype and shared early environmental exposures on the gut microbiome. Moreover, genetically 'identical'[9] monozygotic twin pairs gain weight in response to overfeeding in a more reproducible way than unrelated individuals[10] and are more concordant for body mass index (BMI) than dizygotic twin pairs[11].

Twin pairs who had been enrolled in the Missouri Adolescent Female Twin Study (MOAFTS[12]) were recruited for this study (mean period of enrolment in MOAFTS, 11.7 ± 1.2 years; range, 4.4–13.0 years). Twins were 21–32 years old, of European or African ancestry, and were generally concordant for obesity (BMI $\geq 30\,kg\,m^{-2}$) or leanness (BMI = 18.5–24.9 $kg\,m^{-2}$) (one twin pair was lean/overweight (overweight defined as BMI ≥ 25 and < 30) and six pairs were overweight/obese). They had not taken antibiotics for at least 5.49 ± 0.09 months. Each participant completed a detailed medical, lifestyle and dietary questionnaire: study enrolees were broadly representative of the overall Missouri population for BMI, parity, education and marital status (see Supplementary Results). Although all were born in Missouri, they currently live throughout the USA: 29% live in the same house, but some live more than 800 km apart. Because faecal samples are readily attainable and representative of interpersonal differences in gut microbial ecology[7], they were collected from each individual and frozen immediately. The collection procedure was repeated again with an average interval between sampling of 57 ± 4 days.

To characterize the bacterial lineages present in the faecal microbiotas of these 154 individuals, we performed 16S rRNA sequencing, targeting the full-length gene with an ABI 3730xl capillary sequencer. Additionally, we performed multiplex pyrosequencing with a 454 FLX instrument to survey the gene's V2 variable region[13] and its V6 hypervariable region[14] (Supplementary Tables 1–3).

Complementary phylogenetic and taxon-based methods were used to compare 16S rRNA sequences among faecal communities (see Methods). No matter which region of the gene was examined, individuals from the same family (a twin and her co-twin, or twins and their mother) had a more similar bacterial community structure than unrelated individuals (Fig. 1a and Supplementary Fig. 1a, b), and shared significantly more species-level phylotypes (16S rRNA sequences with ≥97% identity comprise each phylotype) ($G = 55.2$, $P < 10^{-12}$ (V2); $G = 12.3$, $P < 0.001$ (V6); $G = 11.3$, $P < 0.001$ (full-length)). No significant correlation was seen between the degree of physical separation of family members' current homes and the degree of similarity between their microbial communities (defined by UniFrac[15]). The observed familial similarity was not due to an indirect effect of the physiological states of obesity versus leanness; similar results were observed after stratifying twin pairs and their mothers by BMI category (concordant lean or concordant obese individuals; Supplementary Fig. 2). Surprisingly, there was no significant difference in the degree of similarity in the gut microbiotas of adult monozygotic compared with dizygotic twin pairs (Fig. 1a). However, we could not assess whether monozygotic and dizygotic twin pairs had different degrees of similarities at earlier stages of their lives.

Multiplex pyrosequencing of V2 and V6 amplicons allowed higher levels of coverage compared with what was feasible using Sanger sequencing, reaching on average 3,984 ± 232 (V2) and 24,786 ± 1,403 (V6) sequences per sample. To control for differences

[1]Center for Genome Sciences. [2]Department of Psychiatry, Washington University School of Medicine, St Louis, Missouri 63108, USA. [3]Department of Computer Science. [4]Department of Chemistry and Biochemistry, University of Colorado, Boulder, Colorado 80309, USA. [5]CNRS, UMR6098, Marseille, France. [6]Josephine Bay Paul Center, Marine Biological Laboratory, Woods Hole, Massachusetts 02543, USA. [7]Environmental Genomics Core Facility, University of South Carolina, Columbia, South Carolina 29208, USA. [8]Department of Chemistry and Biochemistry and the Advanced Center for Genome Technology, University of Oklahoma, Norman, Oklahoma 73019, USA. [9]454 Life Sciences, Branford, Connecticut 06405, USA.

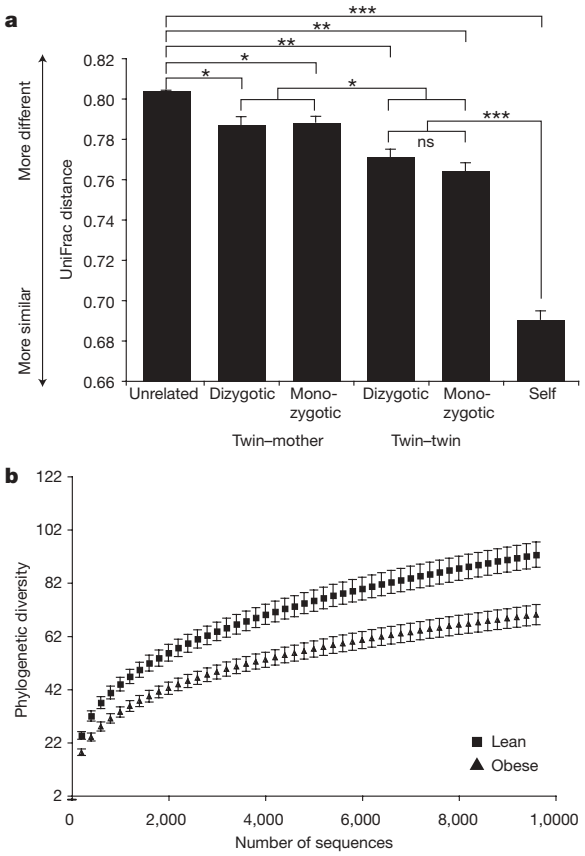


Figure 1 | 16S rRNA gene surveys reveal familial similarity and reduced diversity of the gut microbiota in obese individuals. a, Average unweighted UniFrac distance (a measure of differences in bacterial community structure) between individuals over time (self), twin pairs, twins and their mother, and unrelated individuals (1,000 sequences per V2 data set; Student's *t*-test with Monte Carlo; $*P < 10^{-5}$; $**P < 10^{-14}$; $***P < 10^{-41}$; mean ± s.e.m.). **b**, Phylogenetic diversity curves for the microbiota of lean and obese individuals (based on 1–10,000 sequences per V6 data set; mean ± 95% confidence intervals shown).

in coverage, all analyses were performed on an equal number of randomly selected sequences (200 full-length, 1,000 V2 and 10,000 V6). At this level of coverage, there was little overlap between the sampled faecal communities. Moreover, the number of 16S rRNA gene sequences belonging to each phylotype varied greatly between faecal microbiotas (Supplementary Tables 6–8).

Because this apparent lack of overlap could reflect the level of coverage (Supplementary Tables 1–3), we subsequently searched all hosts for bacterial phylotypes present at high abundance using a sampling model based on a combination of standard Poisson and binomial sampling statistics. The analysis allowed us to conclude that no phylotype was present at more than about 0.5% abundance in all of the samples in this study (see Supplementary Results). Finally, we sub-sampled our data set by randomly selecting 50–3,000 sequences per sample; again, no phylotypes were detectable in all individuals sampled within this range of coverage (Supplementary Fig. 3).

Samples taken from the same individual at the initial collection point and 57 ± 4 days later were consistent with respect to the specific phylotypes found (Supplementary Figs 4 and 5), but showed variations in relative abundance of the major gut bacterial phyla (Supplementary Fig. 6). There was no significant association between UniFrac distance and the time between sample collections. Overall, faecal samples from the same individual were much more similar to one another than samples from family members or unrelated individuals (Fig. 1a), demonstrating that short-term temporal changes in community structure within an individual are minor compared with inter-personal differences.

Analysis of 16S rRNA data sets produced by the three PCR-based methods, plus shotgun sequencing of community DNA (see below), revealed a lower proportion of Bacteroidetes and a higher proportion of Actinobacteria in obese compared with lean individuals of both ancestries (Supplementary Table 9). Combining the individual *P* values across these independent analyses using Fisher's method disclosed significantly fewer Bacteroidetes ($P = 0.003$), more Actinobacteria ($P = 0.002$) but no significant difference in Firmicutes ($P = 0.09$). These findings agree with previous work showing comparable differences in both taxa in mice[2] and a progressive increase in the representation of Bacteroidetes when 12 unrelated, obese humans lost weight after being placed on one of two reduced-calorie diets[6].

Across all methods, obesity was associated with a significant decrease in the level of diversity (Fig. 1b and Supplementary Fig. 1c–f). This reduced diversity suggests an analogy: the obese gut microbiota is not like a rainforest or reef, which are adapted to high energy flux and are highly diverse; rather, it may be more like a fertilizer runoff where a reduced-diversity microbial community blooms with abnormal energy input[16].

We subsequently characterized the microbial lineage and gene content of the faecal microbiomes of 18 individuals representing six of the families (three lean and three obese European ancestry monozygotic twin pairs and their mothers) through shotgun pyrosequencing (Supplementary Tables 4 and 5) and BLASTX comparisons against several databases (KEGG[17] (version 44) and STRING[18]) plus a custom database of 44 reference human gut microbial genomes (Supplementary Figs 7–10 and Supplementary Results). Our analysis parameters were validated using control data sets comprising randomly fragmented microbial genes with annotations in the KEGG database[17] (Supplementary Fig. 11 and Supplementary Methods). We also tested how technical advances that produce longer reads might improve these assignments by sequencing faecal community samples from one twin pair using Titanium pyrosequencing methods (average read length of 341 ± 134 nucleotides (s.d.) versus 208 ± 68 nucleotides for the standard FLX method). Supplementary Fig. 12 shows that the frequency and quality of sequence assignments is improved as read length increases from 200 to 350 nucleotides.

The 18 microbiomes were searched to identify sequences matching domains from experimentally validated carbohydrate-active enzymes (CAZymes). Sequences matching 156 total CAZy families were found within at least one human gut microbiome, including 77 glycoside hydrolase, 21 carbohydrate-binding module, 35 glycosyltransferase, 12 polysaccharide lyase and 11 carbohydrate-esterase families (Supplementary Table 10). On average, 2.62 ± 0.13% of the sequences in the gut microbiome could be assigned to CAZymes (a total of 217,615 sequences), a percentage that is greater than the most abundant KEGG pathway ('Transporters'; 1.20 ± 0.06% of the filtered sequences generated from each sample) and indicative of the abundant and diverse set of microbial genes directed towards accessing a wide range of polysaccharides.

Category-based clustering of the functions from each microbiome was performed using principal components analysis (PCA) and hierarchical clustering[19]. Two distinct clusters of gut microbiomes were identified based on metabolic profile, corresponding to samples with an increased abundance of Firmicutes and Actinobacteria, and samples with a high abundance of Bacteroidetes (Fig. 2a). A linear regression of the first principal component (PC1, explaining 20% of the functional variance) and the relative abundance of the Bacteroidetes showed a highly significant correlation ($R^2 = 0.96$, $P < 10^{-12}$; Fig. 2b). Functional profiles stabilized within each individual's microbiome after 20,000 sequences had been accumulated (Supplementary Fig. 13). Family members had more similar profiles than unrelated individuals (Fig. 2c), suggesting that shared bacterial community structure ('who's there' based on 16S rRNA analyses) also translates into shared community-wide relative abundance of metabolic pathways. Accordingly, a direct comparison of functional

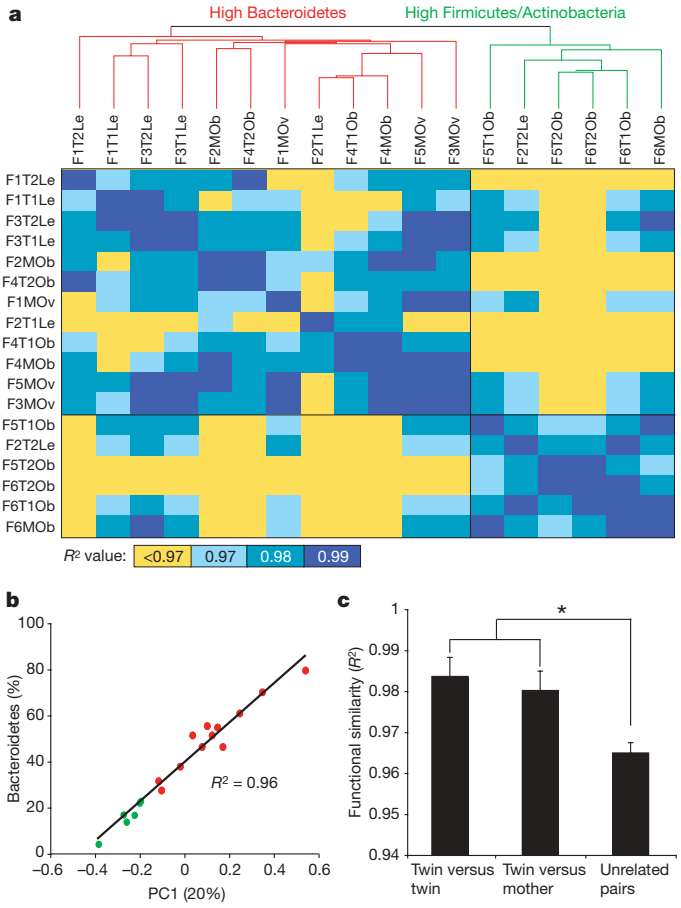


Figure 2 | Metabolic-pathway-based clustering and analysis of the human gut microbiome of monozygotic twins. a, Clustering of functional profiles based on the relative abundance of KEGG metabolic pathways. All pairwise comparisons were made of the profiles by calculating each R^2 value. Sample identifier nomenclature: family number, twin number or mother, and BMI category (Le, lean; Ov, overweight; Ob, obese; for example, F1T1Le stands for family 1, twin 1, lean). **b**, The relative abundance of Bacteroidetes as a function of the first principal component derived from an analysis of KEGG metabolic profiles. **c**, Comparisons of functional similarity between twin pairs, between twins and their mother, and between unrelated individuals. Asterisk indicates significant differences (Student's *t*-test with Monte Carlo; $P < 0.01$; mean ± s.e.m.).

and taxonomic similarity (see Supplementary Methods) disclosed a significant association: individuals with similar taxonomic profiles also share similar metabolic profiles ($P < 0.001$; Mantel test).

Functional clustering of phylum-wide sequence bins representing microbiome reads assigned to 23 human gut Firmicutes and 14 Bacteroidetes reference genomes showed discrete clustering by phylum (Supplementary Figs 14a and 15). Bootstrap analyses of the relative abundance of metabolic pathways in the microbiome-derived Firmicutes and Bacteroidetes sequence bins disclosed 26 pathways with a significantly different relative abundance (Supplementary Fig. 14a). The Bacteroidetes bins were enriched for several carbohydrate metabolism pathways, whereas the Firmicutes bins were enriched for transport systems. This finding is consistent with our CAZyme analysis, which revealed a significantly higher relative abundance of glycoside hydrolases, carbohydrate-binding modules, glycosyltransferases, polysaccharide lyases and carbohydrate esterases in the Bacteroidetes sequence bins (Supplementary Fig. 14b).

One of the major goals of the International Human Microbiome Project(s) is to determine whether there is an identifiable 'core microbiome' of shared organisms, genes or functional capabilities found in a given body habitat of all or the vast majority of humans[1]. Although all of the 18 gut microbiomes surveyed showed a high level of β-diversity with respect to the relative abundance of bacterial phyla (Fig. 3a), analysis of the relative abundance of broad functional categories of genes and metabolic pathways (KEGG) revealed a generally consistent pattern regardless of the sample surveyed (Fig. 3b and Supplementary Table 11): the pattern is also consistent with results we obtained from a meta-analysis of previously published gut microbiome data sets from nine adults[20,21] (Supplementary Fig. 16). This consistency is not simply due to the broad level of these annotations, as a similar analysis of Bacteroidetes and Firmicutes reference genomes revealed substantial variation in the relative abundance of each category (see Supplementary Fig. 17). Furthermore, pairwise comparisons of metabolic profiles obtained from the 18 microbiomes in this study revealed an average value of R^2 of 0.97 ± 0.002 (Fig. 2a), indicating a high level of functional similarity.

Overall functional diversity was compared using the Shannon index[22], a measurement that combines diversity (the number of different metabolic pathways) and evenness (the relative abundance of each pathway). The human gut microbiomes surveyed had a stable and high Shannon index value (4.63 ± 0.01), close to the maximum possible level of functional diversity (5.54; see Supplementary Methods). Despite the presence of a small number of abundant metabolic pathways (listed in Supplementary Table 11), the overall functional profile of each gut microbiome is quite even (Shannon evenness of 0.84 ± 0.001 on a scale of 0–1), demonstrating that most metabolic pathways are found at a similar level of abundance. Interestingly, the level of functional diversity in each microbiome was significantly linked to the relative abundance of the Bacteroidetes ($R^2 = 0.81$, $P < 10^{-6}$); microbiomes enriched for Firmicutes/Actinobacteria had a lower level of functional diversity. This observation is consistent with an analysis of simulated metagenomic reads generated from each of 36 Bacteroidetes and Firmicutes genomes (Supplementary Fig. 18): on average, the Bacteroidetes genomes have a significantly higher level of both functional diversity and evenness (Mann–Whitney *U*-test, $P < 0.01$).

At a finer level, 26–53% of 'enzyme'-level functional groups (KEGG/CAZy/STRING) were shared across all 18 microbiomes, whereas 8–22% of the groups were unique to a single microbiome (Supplementary Fig. 19a–c). The 'core' functional groups present in all microbiomes were also highly abundant, representing 93–98% of the total sequences. Given the higher relative abundance of these 'core' groups, more than 95% were found after 26.11 ± 2.02 megabases of sequence were collected from a given microbiome, whereas the 'variable' groups continued to increase substantially with each additional megabase of sequence. Of course, any estimate of the total size of the core microbiome will depend on sequencing effort, especially for

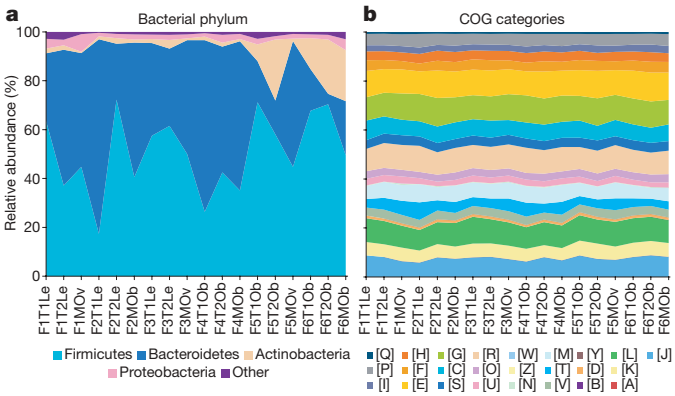


Figure 3 | Comparison of taxonomic and functional variations in the human gut microbiome. a, Relative abundance of major phyla across 18 faecal microbiomes from monozygotic twins and their mothers, based on BLASTX comparisons of microbiomes and the National Center for Biotechnology Information non-redundant database. **b**, Relative abundance of categories of genes across each sampled gut microbiome (letters correspond to categories in the COG database).

functional groups found at a low abundance. On average, our survey achieved more than 450,000 sequences per faecal sample, which, assuming an even distribution, would allow us to sample groups found at a relative abundance of 10^{-4}. To estimate the total size of the core microbiome based on the 18 individuals, we randomly subsampled each microbiome in 1,000 sequence intervals (Supplementary Fig. 19d). Based on this analysis, the core microbiome is approaching a total of 2,142 total orthologous groups (one site binding (hyperbola) curve fit, $R^2 = 0.9966$), indicating that we identified 93% of functional groups (defined by STRING) found within the core microbiome of the 18 individuals surveyed. Of these core groups, 71% (CAZy), 64% (KEGG) and 56% (STRING) were also found in the nine previously published, but much lower coverage, data sets generated by capillary sequencing of adult faecal DNA[20,21] (average of 78,413 ± 2,044 bidirectional reads per sample; see Supplementary Methods).

Metabolic reconstructions of the 'core' microbiome revealed significant enrichment for several expected functional categories, including those involved in transcription and translation (Fig. 4). Metabolic profile-based clustering indicated that the representation of 'core' functional groups was highly consistent across samples (Supplementary Fig. 20), and included several pathways that are likely important for life in the gut, such as those for carbohydrate and amino-acid metabolism (for example, fructose/mannose metabolism, amino-sugar metabolism and N-glycan degradation). Variably represented pathways and categories include cell motility (only a subset of Firmicutes produce flagella), secretion systems and membrane transport (for example, phosphotransferase systems involved in the import of nutrients, including sugars; Fig. 4 and Supplementary Fig. 20).

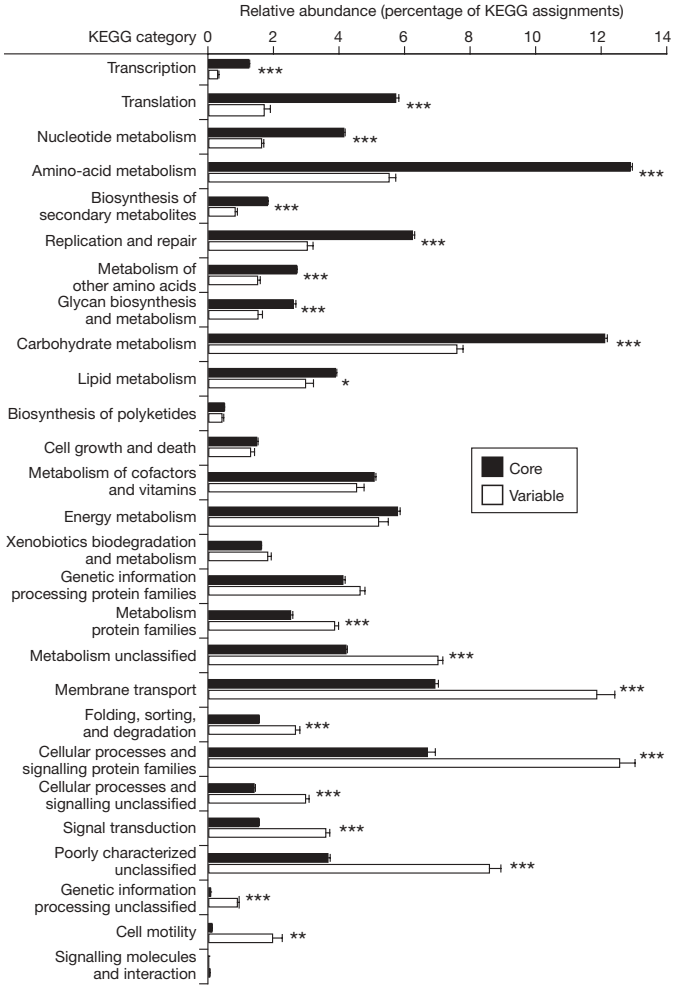


Figure 4 | KEGG categories enriched or depleted in the core versus variable components of the gut microbiome. Sequences from each of the 18 faecal microbiomes were binned into the 'core' or 'variable' microbiome based on the co-occurrence of KEGG orthologous groups (core groups were found in all 18 microbiomes whereas variable groups were present in fewer (<18) microbiomes; see Supplementary Fig. 19a). Asterisks indicate significant differences (Student's *t*-test, $*P < 0.05$, $**P < 0.001$, $***P < 10^{-5}$; mean ± s.e.m.).

To identify metabolic pathways associated with obesity, only non-core associated (variable) functional groups were included in a comparison of the gut microbiomes of lean versus obese twin pairs. A bootstrap analysis[23] was used to identify metabolic pathways that were enriched or depleted in the variable obese gut microbiome. For example, similar to a mouse model of diet-induced obesity[4], the obese human gut microbiome was enriched for phosphotransferase systems involved in microbial processing of carbohydrates (Supplementary Table 12). All gut microbiome sequences were compared with the custom database of 44 human gut genomes: an odds ratio analysis revealed 383 genes that were significantly different between the obese and lean gut microbiome (q value < 0.05; 273 enriched and 110 depleted in the obese microbiome; Supplementary Tables 13 and 14). By contrast, only 49 genes were consistently enriched or depleted between all twin pairs (see Supplementary Methods).

These obesity-associated genes were representative of the taxonomic differences described above: 75% of the obesity-enriched genes were from Actinobacteria (compared with 0% of lean-enriched genes; the other 25% are from Firmicutes) whereas 42% of the lean-enriched genes were from Bacteroidetes (compared with 0% of the obesity-enriched genes). Their functional annotation indicated that many are involved in carbohydrate, lipid and amino-acid metabolism (Supplementary Tables 13 and 14). Together, they comprise an initial set of microbial biomarkers of the obese gut microbiome.

Our finding that the gut microbial community structures of adult monozygotic twin pairs had a degree of similarity that was comparable to that of dizygotic twin pairs, and only slightly more similar than that of their mothers, is consistent with an earlier fingerprinting study of adult twins[24], and with a recent microarray-based analysis, which revealed that gut community assembly during the first year of life followed a more similar pattern in a pair of dizygotic twins than 12 unrelated infants[25]. Intriguingly, another fingerprinting study of monozygotic and dizygotic twins in childhood showed a slightly reduced similarity profile in dizygotic twins[26]. Thus, comprehensive time-course studies, comparing monozygotic and dizygotic twin pairs from birth through adulthood, as well as intergenerational analyses of their families' microbiotas, will be key to determining the relative contributions of host genotype and environmental exposures to (gut) microbial ecology.

The hypothesis that there is a core human gut microbiome, definable by a set of abundant microbial organismal lineages that we all share, may be incorrect: by adulthood, no single bacterial phylotype was detectable at an abundant frequency in the guts of all 154 sampled humans. Instead, it appears that a core gut microbiome exists at the level of shared genes, including an important component involved in various metabolic functions. This conservation suggests a high degree of redundancy in the gut microbiome and supports an ecological view of each individual as an 'island' inhabited by unique

collections of microbial phylotypes: as in actual islands, different species assemblages converge on shared core functions provided by distinctive components. Our findings raise the question of how core functionality is assembled in this body habitat. Understanding the underlying principles should provide insights about microbial adaptation to, and mutualistic community assembly within, a wide range of environments.

METHODS SUMMARY

Faecal samples were collected from each individual. Community DNA was prepared and used for pyrosequencing (454 Life Sciences), as well as for PCR and sequencing of bacterial 16S rRNA genes. Shotgun reads were mapped to reference genomes using the National Center for Biotechnology Information 'non-redundant' database, KEGG[17], STRING[18], CAZy (http://www.cazy.org/) and a 44-member human-gut microbial genome database. Metabolic reconstructions were performed based on CAZy, KEGG and STRING annotations. The relative abundance of KEGG metabolic pathways is referred to as a 'metabolic profile'.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 29 June; accepted 14 October 2008.
Published online 30 November 2008.

1. Turnbaugh, P. J. *et al.* The human microbiome project. *Nature* **449,** 804–810 (2007).
2. Ley, R. E. *et al.* Obesity alters gut microbial ecology. *Proc. Natl Acad. Sci. USA* **102,** 11070–11075 (2005).
3. Turnbaugh, P. J. *et al.* An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444,** 1027–1031 (2006).
4. Turnbaugh, P. J., Bäckhed, F., Fulton, L. & Gordon, J. I. Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome. *Cell Host Microbe* **3,** 213–223 (2008).
5. Bäckhed, F. *et al.* The gut microbiota as an environmental factor that regulates fat storage. *Proc. Natl Acad. Sci. USA* **101,** 15718–15723 (2004).
6. Ley, R. E., Turnbaugh, P. J., Klein, S. & Gordon, J. I. Human gut microbes associated with obesity. *Nature* **444,** 1022–1023 (2006).
7. Eckburg, P. B. *et al.* Diversity of the human intestinal microbial flora. *Science* **308,** 1635–1638 (2005).
8. Frank, D. N. *et al.* Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc. Natl Acad. Sci. USA* **104,** 13780–13785 (2007).
9. Bruder, C. E. *et al.* Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. *Am. J. Hum. Genet.* **82,** 763–771 (2008).
10. Bouchard, C. *et al.* The response to long-term overfeeding in identical twins. *N. Engl. J. Med.* **322,** 1477–1482 (1990).
11. Maes, H. H., Neale, M. C. & Eaves, L. J. Genetic and environmental factors in relative body weight and human adiposity. *Behav. Genet.* **27,** 325–351 (1997).
12. Heath, A. C. *et al.* Ascertainment of a mid-western US female adolescent twin cohort for alcohol studies: assessment of sample representativeness using birth record data. *Twin Res.* **5,** 107–112 (2002).
13. Hamady, M., Walker, J. J., Harris, J. K., Gold, N. J. & Knight, R. Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nature Methods* **5,** 235–237 (2008).
14. Sogin, M. L. *et al.* Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc. Natl Acad. Sci. USA* **103,** 12115–12120 (2006).
15. Lozupone, C., Hamady, M. & Knight, R. UniFrac-an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics* **7,** 371 (2006).
16. Watson, S. B., McCauley, E. & Downing, J. A. Patterns in phytoplankton taxonomic composition across temperate lakes of differing nutrient status. *Limnol. Oceanogr.* **42,** 487–495 (1997).
17. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32,** D277–D280 (2004).
18. von Mering, C. *et al.* STRING 7-recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.* **35,** D358–D362 (2007).
19. de Hoon, M. J., Imoto, S., Nolan, J. & Miyano, S. Open source clustering software. *Bioinformatics* **20,** 1453–1454 (2004).
20. Gill, S. R. *et al.* Metagenomic analysis of the human distal gut microbiome. *Science* **312,** 1355–1359 (2006).
21. Kurokawa, K. *et al.* Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res.* **14,** 169–181 (2007).
22. Dinsdale, E. A. *et al.* Functional metagenomic profiling of nine biomes. *Nature* **452,** 629–632 (2008).
23. Rodriguez-Brito, B., Rohwer, F. & Edwards, R. A. An application of statistics to comparative metagenomics. *BMC Bioinformatics* **7,** 162 (2006).
24. Zoetandal, E. G., Akkermans, A. D. L., Akkermans-van Vliet, W. M., de Visser, J. A. & de Vos, W. M. The host genotype affects the bacterial community in the human gastrointestinal tract. *Microb. Ecol. Health Dis.* **13,** 129–134 (2001).
25. Palmer, C., Bik, E. M., Digiulio, D. B., Relman, D. A. & Brown, P. O. Development of the human infant intenstinal microbiota. *PLoS Biol.* **5,** e177 (2007).
26. Stewart, J. A., Chadwick, V. S. & Murray, A. Investigations into the influence of host genetics on the predominant eubacteria in the faecal microflora of children. *J. Med. Microbiol.* **54,** 1239–1242 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank: S. Wagoner and J. Manchester for technical support; S. Marion and D. Hopper for recruitment of participants and sample collection; A. Goodman, B. Muegge, and M. Mahowald for suggestions; S. Huse (Marine Biological Laboratory), F. Niazi and S. Attiya (454 Life Sciences), C. Markovic, L. Fulton, B. Fulton, E. Mardis and R. Wilson (Washington University Genome Sequencing Center) and S. Macmil, G. Wiley, C. Qu, and P. Wang (University of Oklahoma) for their assistance with sequencing; and P. M. Coutinho (Université de Provence, France) for help with the CAZy analysis. Deep draft assemblies of reference gut genomes were generated as part of a National Human Genome Research Institute (NHGRI)-sponsored human gut microbiome initiative (http://genome.wustl.edu/pub/organism/Microbes/Human_Gut_Microbiome/). This work was supported in part by the National Institutes of Health (DK78669/ES012742/AA09022/HD049024), the National Science Foundation (OCE0430724), the W.M. Keck Foundation, and the Crohn's and Colitis Foundation of America.

Author Contributions P.J.T., A.C.H., R.K. and J.I.G. designed the experiments. P.J.T., T.Y., A.D., R.E.L., M.L.S., W.J.J., B.A.R., J.P.A. and M.E. generated the data. P.J.T., M.H., M.L.S., B.L.C., A.D., B.H., A.C.H., R.K. and J.I.G. analysed the data. P.J.T., A.C.H., R.K. and J.I.G. wrote the manuscript with input from the other members of the team.

Author Information This Whole Genome Shotgun project is deposited in DDBJ/EMBL/GenBank under accession number 32089. 454 pyrosequencing reads are deposited in the NCBI Short Read Archive. Nearly full-length 16S rRNA gene sequences are deposited in GenBank under accession numbers FJ362604–FJ372382. Annotated sequences are also available in MG-RAST (http://metagenomics.nmpdr.org/). 454-generated 16S rRNA sequences with sample identifiers are also available at http://gordonlab.wustl.edu/SuppData.html. Correspondence and requests for materials should be addressed to J.I.G. (jgordon@wustl.edu).

METHODS

Community DNA preparation. Faecal samples were frozen immediately after they were produced. De-identified samples were stored at −80 °C before processing. Ten to twenty grams of each sample was pulverized in liquid nitrogen with a mortar and pestle. An aliquot (approximately 500 mg) of each sample was then suspended, while frozen, in a solution containing 500 µl of extraction buffer (200 mM Tris (pH 8.0), 200 mM NaCl, 20 mM EDTA), 210 µl of 20% SDS, 500 µl of a mixture of phenol:chloroform:isoamyl alcohol (25:24:1, pH 7.9), and 500 µl of a slurry of 0.1-mm diameter zirconia/silica beads (BioSpec Products). Microbial cells were subsequently lysed by mechanical disruption with a bead beater (BioSpec Products) set on high for 2 min at room temperature, followed by extraction with phenol:chloroform:isoamyl alcohol, and precipitation with isopropanol. DNA obtained from three separate aliquots of each faecal sample were pooled (≥200 µg DNA) and used for pyrosequencing (see below).

16S rRNA gene-sequence-based surveys. Complementary phylogenetic- and taxon-based methods were used to compare 16S rRNA sequences among faecal communities. Phylogenetic clustering with UniFrac[15] is based on the principle that communities can be compared in terms of their shared evolutionary history, as measured by the degree to which they share branch length on a phylogenetic tree. We complemented this approach with taxon-based methods[27], which disregard some of the information contained in the phylogenetic tree of the taxa in question, but have the advantage that specific taxa unique to, or shared among, groups of samples can be identified (for example, those from lean or obese individuals). Before both types of analysis, we grouped 16S rRNA gene sequences into operational taxonomic units (OTUs/phylotypes) using both cd-hit[28] and the furthest-neighbour-like algorithm, with a sequence identity threshold of 97%, which is commonly used to define 'species'-level phylotypes. Taxonomy was assigned using the best-BLAST-hit against Greengenes[29] (E value cutoff of 10^{-10}, minimum 88% coverage, 88% identity) and the Hugenholtz taxonomy (downloaded from http://greengenes.lbl.gov/Download/Sequence_Data/Greengenes_format/ on 12 May 2008, excluding sequences annotated as chimaeric).

Selection of operational taxonomic units. 16S rRNA gene-derived pyrosequencing data were pre-processed to remove sequences with low-quality scores, sequences with ambiguous characters or sequences outside the length bounds (V6 < 50 nucleotides, V2 < 200 nucleotides), and binned according to sample-specific barcode (see, for example, ref. 13). Similar sequences were identified using Megablast[30] and cd-hit, with the following parameters: E value 10^{-10} (Megablast only); minimum coverage 99%; minimum pairwise identity 97%. Candidate OTUs were identified as sets of sequences connected to each other at this level using a maximum of 4,000 hits per sequence. Each candidate OTU was considered valid if the average density of connection was above threshold; otherwise, it was broken up into smaller connected components[27].

Tree building and UniFrac clustering for PCA analysis. A relaxed neighbour-joining tree was built from one representative sequence per OTU using Clearcut[31], employing the Kimura correction (the PH Lane mask was applied to V2 and full-length data), but otherwise with default comparisons. Unweighted UniFrac[15] was run using the resulting tree. PCA was performed on the resulting matrix of distances between each pair of samples. To determine if the UniFrac distances were on average significantly different for pairs of samples (that is, between twin pairs, between twins and their mother, or between unrelated individuals), we performed a t-test on the UniFrac distance matrix, and generated a P value for the t-statistic by permutation of the rows and columns as in the Mantel test, regenerating the t-statistic for 1,000 random samples, and using the distribution to obtain an empirical P value.

Rarefaction and phylogenetic diversity measurements. To determine which individuals had the most diverse communities of gut bacteria, rarefaction plots and phylogenetic diversity measurements, as described by Faith[32], were made for each sample. Phylogenetic diversity is the total amount of branch length in a phylogenetic tree constructed from the combined 16S rRNA data sets, leading to the sequences in a given sample. To account for differences in sampling effort between individuals, and to estimate how far we were from sampling the diversity of each individual completely, we plotted the accumulation of phylogenetic diversity (branch length) with sampling effort, in a manner analogous to rarefaction curves. We generated the phylogenetic diversity rarefaction curve for each individual by applying custom python code (http://bmf2.colorado.edu/unifrac/about.psp) to the Arb parsimony insertion tree[27].

Pyrosequencing of total community DNA. Shotgun sequencing runs were performed on the 454 FLX pyrosequencer from total faecal community DNA. Two samples were also analysed in a single run using Titanium extra-long-read pyrosequencing technology (see Supplementary Tables 4 and 5). Sequencing reads with degenerate bases ('Ns') were removed along with all duplicate sequences, as sequences of identical length and content are a common artefact of the pyrosequencing methodology. Finally, human sequences were removed by identifying sequences homologous to the *Homo sapiens* reference genome (BLASTN $E < 10^{-5}$, %identity > 75, score > 50).

CAZyme analysis. Metagenomic sequence reads were searched against a library of modules derived from all entries in the carbohydrate-active enzymes (CAZy) database (www.cazy.org using FASTY[33], $E < 10^{-6}$). This library consists of approximately 180,000 previously annotated modules (catalytic modules, carbohydrate-binding modules and other non-catalytic modules or domains of unknown function) derived from about 80,000 protein sequences. The number of sequencing reads matching each CAZy family was divided by the number of total sequences assigned to CAZymes and multiplied by 100 to calculate a relative abundance. An R^2 value was calculated for each pair of CAZy profiles. We then compared the distribution of glycoside hydrolase similarity scores with the distribution of glycosyltransferase similarity scores.

Statistical analyses. Xipe[23] (version 2.4) was used for bootstrap analyses of pathway enrichment and depletion, using the parameters sample size = 10,000 and confidence level = 0.95. Linear regressions were performed in Excel (version 11.0, Microsoft). Mann–Whitney and Student's t-tests were used to identify statistically significant differences between two groups (Prism version 4.0, GraphPad; Excel version 11.0, Microsoft). The Bonferroni correction was used to correct for multiple hypotheses. The Mantel test was used to compare distance matrices: the matrix of each pairwise comparison of the abundance of each reference genome, and the abundance of each metabolic pathway, were compared (Mantel program in Python using PyCogent[34]; 10,000 replicates). Data are represented as mean ± s.e.m. unless otherwise indicated.

Microbiome sequences were compared against the custom database of 44 gut genomes (BLASTX $E < 10^{-5}$, bitscore > 50, and %identity > 50). A gene-by-sample matrix was then screened to identify genes 'commonly-enriched' in either the obese or lean gut microbiome (defined by an odds ratio greater than 2 or less than 0.5 when comparing the pooled obese twin microbiomes with the pooled lean twin microbiomes, and when comparing each individual obese twin microbiome with the aggregate lean twin microbiome, or vice versa). The statistical significance of enriched or depleted genes was then calculated using a modified t-test (q value < 0.05; calculated with code supplied by M. Pop and J.R. White, University of Maryland). We also searched for genes that were consistently enriched or depleted in all six monozygotic twin pairs. A gene-by-sample matrix was generated based on BLASTX comparisons of each microbiome with our custom 44-genome database, to calculate an odds ratio based on the frequency of each gene in each twin versus the respective co-twin. The analysis revealed only 49 genes (odds ratio > 2 or < 0.5): they represent a variety of taxonomic groups, including Firmicutes, Bacteroidetes and Actinobacteria, and did not show any clear functional trends.

27. Ley, R. E. *et al.* Evolution of mammals and their gut microbes. *Science* **320,** 1647–1651 (2008).
28. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22,** 1658–1659 (2006).
29. DeSantis, T. Z. *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72,** 5069–5072 (2006).
30. Zhang, Z., Schwartz, S., Wagner, L. & Miller, W. A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* **7,** 203–214 (2000).
31. Sheneman, L., Evans, J. & Foster, J. A. Clearcut: a fast implementation of relaxed neighbor joining. *Bioinformatics* **22,** 2823–2824 (2006).
32. Faith, D. P. Conservation evaluation and phylogenetic diversity. *Biol. Conserv.* **61,** 1–10 (1992).
33. Pearson, W. R., Wood, T., Zhang, Z. & Miller, W. Comparison of DNA sequences with protein sequences. *Genomics* **46,** 24–36 (1997).
34. Knight, R. *et al.* PyCogent: a toolkit for making sense from sequence. *Genome Biol.* **8,** R171 (2007).

LETTERS

Protein kinase R reveals an evolutionary model for defeating viral mimicry

Nels C. Elde[1], Stephanie J. Child[2], Adam P. Geballe[2,3,4] & Harmit S. Malik[1]

Distinguishing self from non-self is a fundamental biological challenge. Many pathogens exploit the challenge of self discrimination by employing mimicry to subvert key cellular processes including the cell cycle, apoptosis and cytoskeletal dynamics[1–5]. Other mimics interfere with immunity[6,7]. Poxviruses encode K3L, a mimic of eIF2α, which is the substrate of protein kinase R (PKR), an important component of innate immunity in vertebrates[8,9]. The PKR–K3L interaction exemplifies the conundrum imposed by viral mimicry. To be effective, PKR must recognize a conserved substrate (eIF2α) while avoiding rapidly evolving substrate mimics such as K3L. Using the PKR–K3L system and a combination of phylogenetic and functional analyses, we uncover evolutionary strategies by which host proteins can overcome mimicry. We find that PKR has evolved under intense episodes of positive selection in primates. The ability of PKR to evade viral mimics is partly due to positive selection at sites most intimately involved in eIF2α recognition. We also find that adaptive changes on multiple surfaces of PKR produce combinations of substitutions that increase the odds of defeating mimicry. Thus, although it can seem that pathogens gain insurmountable advantages by mimicking cellular components, host factors such as PKR can compete in molecular 'arms races' with mimics because of evolutionary flexibility at protein interaction interfaces challenged by mimicry.

To counteract viral infections, PKR phosphorylates the translation initiation factor eIF2α in the presence of double-stranded RNA (dsRNA) from viruses[8,9]. This activity strongly inhibits protein synthesis and blocks the production of new virus particles. The crucial function of PKR in innate immunity is reflected by the evolution of numerous factors from various viruses that disable PKR to promote viral production[10], including a poxvirus-encoded mimic of eIF2α called K3L (Supplementary Fig. 1). Host proteins such as PKR, which interact directly with viral antagonists such as K3L, can be subject to molecular 'arms-races' in which amino-acid substitutions that directly affect interactions can be rapidly fixed by positive selection[11,12].

To determine whether PKR might be subject to positive selection, we cloned and sequenced complementary DNA of PKR from a panel of 20 primates representing more than 30 million years of evolutionary divergence. By considering ratios of the rates of non-synonymous (dN) and synonymous (dS) substitutions, we found evidence for ancient, episodic positive selection in primate lineages ($P < 0.0003$; Fig. 1a and Supplementary Table 1). In particular, one branch in Old World monkeys was calculated to have undergone 22 non-synonymous substitutions without any synonymous changes, one of the most intense episodes of positive selection reported for any primate gene (Supplementary Data). Likelihood ratio tests[13] using the entire phylogeny reveal that 17% of codons evolved with an average dN/dS ratio of 3.7, strongly supporting a finding of positive selection ($P < 0.0001$; Supplementary Tables 2 and 3), even after accounting for the potentially confounding effects of recombination and synonymous site variation[14] ($P < 0.0001$; Supplementary Tables 4 and 5). Positive selection is observed in each of the three domains of PKR—the dsRNA-binding domain, the spacer region and even the kinase domain—which is consistent with an extensive history of facing viral factors that directly bind PKR in these separate domains (Supplementary Fig. 1). Several residues in the kinase domain, which make direct contacts with eIF2α (ref. 15), are among the fastest-evolving residues in PKR (Fig. 1b and Supplementary Fig. 1), suggesting that selective pressure to evade eIF2α mimics may have driven changes in these residues.

Similarly, we find that positive selection has acted on the eIF2α mimic K3L (Supplementary Fig. 2). For instance, in a comparison of K3L from variola major (smallpox) and vaccinia viruses, we find a dN/dS ratio of 2.80 ($P < 0.001$), whereas fewer than 10% of orthologues in vaccinia and variola comparisons show any evidence of positive selection (average dN/dS = 0.10; N.C.E. and H.S.M., unpublished observations). This suggests that poxviral eIF2α mimics have also undergone positive selection and reflects the possibility that K3L has not achieved or maintained an optimal state of mimicry. Instead, K3L might continually evolve to counter adaptive changes in PKR.

In contrast with the rapid evolution of PKR, its substrate, eIF2α, is essentially unchanged in simian primates at the amino-acid level (dN/dS = 0 in a comparison of human and rhesus). Thus, PKR must recognize an unchanging substrate while evolving to discriminate against mimics such as K3L to be effective. If we consider that most viruses, including poxviruses[16,17], evolve at faster rates than primates, such challenges by mimics are daunting for hosts. Nevertheless, PKR can inhibit viruses encoding eIF2α mimics[10], suggesting that adaptive changes in PKR might help to overcome mimicry by these factors.

We investigated whether primate PKR orthologues differ in their ability to discriminate against K3L from vaccinia, the model poxvirus. Because the entire clade of extant poxviruses is very young relative to the divergence between primates[16,17], we could not investigate strict co-evolutionary dynamics between PKR and K3L. Instead, we used vaccinia K3L as a means to study the evolutionary strategies afforded PKR for counteracting substrate mimics that were faced over the course of primate evolution, which could leave PKR variants either susceptible or resistant to vaccinia K3L. Even though primate *PKR* alleles did not necessarily evolve under pressure from vaccinia K3L, our approach allowed us to identify the mechanisms by which host proteins might defeat mimicry more generally. Examining host–virus evolution from a similar perspective led to the identification of a region in the restriction factor Trim5α that confers specificity against ancient, extinct retroviruses but fails to protect humans from HIV[18,19].

A growth assay in yeast has provided a simple test of PKR function[20]. Human PKR recognizes and phosphorylates yeast eIF2α, as a result of its high level of similarity to primate eIF2α, to cause growth

[1]Division of Basic Sciences, [2]Division of Human Biology, and [3]Division of Clinical Research, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA. [4]Departments of Medicine and Microbiology, University of Washington, Seattle, Washington 98115, USA.

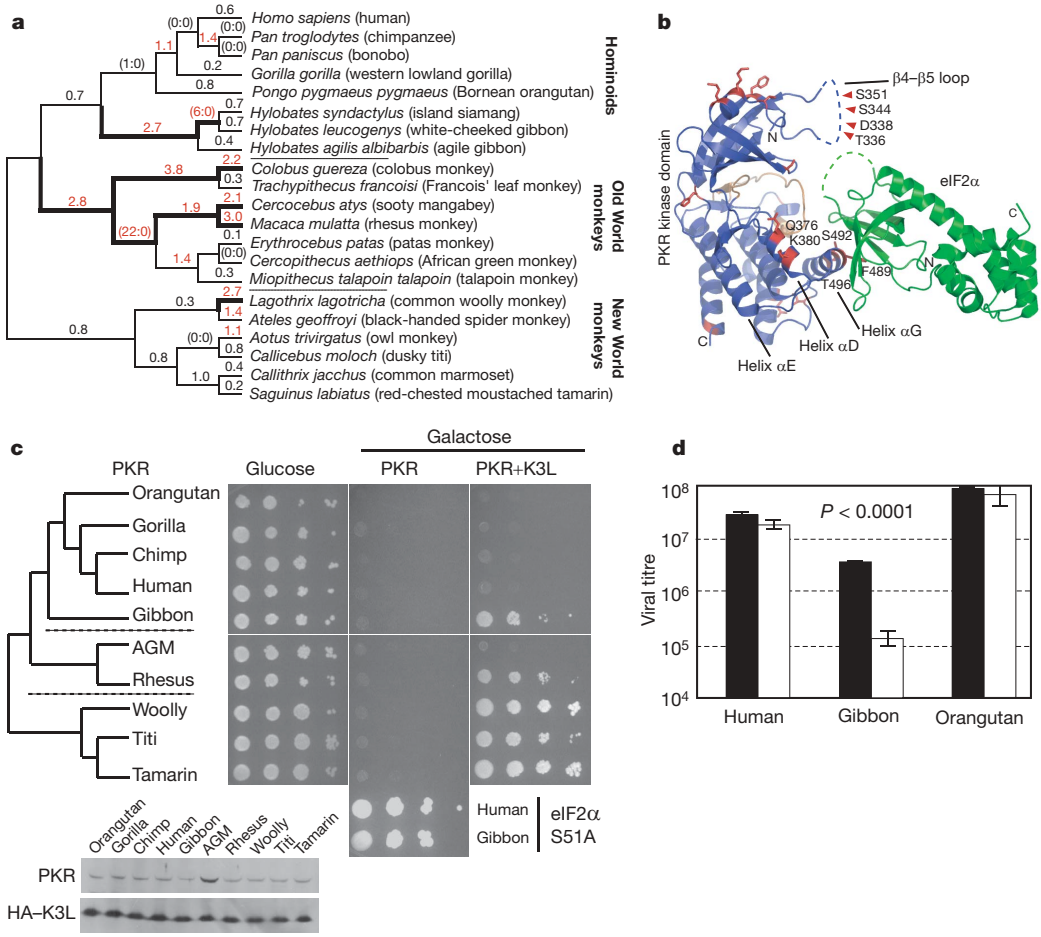


Figure 1 | Widespread positive selection has shaped PKR throughout primate evolution. **a**, *PKR* was sequenced from simian primates that together represent more than 30 million years of divergence. d*N*/d*S* values along each branch of the phylogeny are listed, and those with d*N*/d*S* > 1 are highlighted in red. Branches with bold lines, overlapping the set in red, indicate lineages found to be under positive selection by complementary model fitting analysis (see also Supplementary Table 6). Values in parentheses are shown for branches where no synonymous changes were observed ($S = 0$) and indicate the number of non-synonymous changes (N). **b**, Sites under positive selection (red) are mapped onto a ribbons representation of the complex of the PKR kinase domain (blue) with eIF2α (green) (PDB code 2A1A)[15]. The active site of PKR is shown in orange, and for technical reasons a large portion of the β4–β5 loop (dashed blue line) is invisible from the structure deduced from the co-crystal[15]. Residues under positive selection near the interface of PKR with eIF2α and K3L are noted in the β4–β5 loop (Thr 336, Asp 338, Ser 344, Ser 351) and the αD (Gln 376, Lys 380) and αG (Phe 489, Ser 492, Thr 496) helices. **c**, Plasmids encoding PKR variants from a panel of primates under pGal were introduced into yeast strains HM3 (eIF2α), HM2 (eIF2α and haemagglutinin (HA)-epitope-tagged vaccinia K3L) and J223 (eIF2α-S51A). Tenfold serial dilutions of transformants were spotted on plates containing either glucose or galactose (see Methods). Immunoblot analysis of PKR (top panel) and HA–K3L (bottom panel) is also shown (see Methods). For African green monkey (AGM), resistance to K3L might reflect differences in PKR expression in yeast. **d**, Primary fibroblasts from the indicated primates were infected in triplicate with wild-type (filled columns) or ΔK3L (open columns) vaccinia virus (multiplicity of infection 0.001). Virus production was assessed three days after infection by titring cell lysates. The significance of wild-type virus compared with ΔK3L is indicated (Student's *t*-test; error bars show s.d.). Minor variations of this experiment (not shown) revealed that ΔK3L infections typically produced about fivefold less virus than wild-type virus in gibbon cells.

arrest[15,21]. We expressed ten divergent primate *PKR* cDNAs in yeast to determine whether they differed in their ability to phosphorylate eIF2α. All primate *PKR* genes tested caused consistent levels of growth arrest, which specifically depended on phosphorylation of eIF2α (ref. 22) (Fig. 1c, middle). However, co-expression with vaccinia virus K3L uncovered marked differences in K3L inhibition of primate PKR orthologues, which leads to a rescue of growth[23] (Fig. 1c, right). *PKR* alleles from Old World and New World monkeys, and from white-cheeked gibbon, were generally quite susceptible to suppression of growth arrest by K3L from vaccinia and variola, whereas other hominoid *PKR* alleles showed only modest suppression by K3L (Fig. 1c and Supplementary Fig. 3). Thus, rapid evolution of primate PKR did not seem to alter eIF2α recognition significantly, but resulted in considerable differences in susceptibility to K3L. In particular, we find in the hominoid lineage that human, chimp, gorilla and orangutan PKR orthologues are 1,000-fold more resistant than gibbon PKR to growth rescue by K3L.

We further corroborated the large differences in K3L susceptibility uncovered by the yeast assay by infecting human, orangutan and gibbon fibroblast cell lines with either wild-type vaccinia virus or a strain with a *K3L* gene deletion (ΔK3L). Consistent with our yeast assays and previous reports in human cells[24] was our finding that ΔK3L virus had no significant effect on viral titre in human or orangutan cells but led to a substantial decrease in titre in gibbon cells (Fig. 1d). Vaccinia virus therefore depends on K3L for full infectivity in gibbon cells, where PKR is susceptible to K3L.

We wished to map critical genetic differences between 'resistant' and 'susceptible' *PKR* alleles to understand the basis of resistance to K3L. We first investigated helix αG of the kinase domain because residues 489, 492 and 496 have key functions in the recognition of eIF2α (ref. 15), yet they have evolved under recurrent positive selection (Figs 1b and 2). Whereas gibbon PKR (helix αG: Tyr 489–Ala 492–Thr 496 or Y-A-T) is susceptible to K3L in growth assays, the human αG configuration (F-S-T) in an otherwise gibbon PKR

backbone increases the resistance of gibbon PKR to vaccinia K3L (Fig. 2a, rows 1 and 2). In fact, the A492S substitution (Y-S-T) alone confers on gibbon PKR greatly increased resistance to K3L (Fig. 2a, row 3). These findings reveal that even a single change in PKR at the common interface with substrate and mimic has the capacity to reverse a 'susceptibility' phenotype.

A second determinant, not in helix αG, explains the resistance of orangutan PKR to K3L. When we tested the αG configuration (S-A-K) of the 'resistant' orangutan *PKR* allele (Fig. 1c) in the gibbon backbone, this S-A-K allele was still quite susceptible (Fig. 2a, row 5). To identify the source of resistance of orangutan PKR, we tested chimaeras between orangutan and gibbon PKR and found that a region in the kinase domain containing helices αD and αE from orangutan PKR greatly increased the resistance of gibbon PKR to K3L (data not shown). When we tested individual substitutions in this region, we found that the F394L substitution of the αE helix conferred gibbon PKR with resistance against K3L (Fig. 2c). Importantly, the opposite L394F substitution greatly reduced resistance in orangutan PKR (Fig. 2c). Unlike helix αG, helix αE discrimination seems to be independent of PKR contact with its substrate because it is positioned away from the eIF2α interface (Fig. 2d)[15]. In addition, positive selection in helix αD suggested that this region could contribute to escaping mimicry, either directly or by virtue of co-evolution between helices αD and αG (Supplementary Table 9)[25]. However, we did not find functional evidence for the involvement of αD in resisting K3L over the evolutionary timeframe that we examined for this particular mimic (Supplementary Fig. 4). Therefore, susceptible gibbon *PKR* alleles can gain resistance to vaccinia K3L by single substitutions in either the αG helix or the αE helix (Fig. 2e), increasing the chances of escaping mimicry.

Our analyses suggested that human PKR contained residues associated with increased resistance to K3L from both αG and αE helices. Indeed, we found that a human *PKR* allele carrying 'susceptible' mutations in both its αE (L394F) and αG (F489Y/S492A) helices loses wild-type resistance to K3L (Fig. 3a, row 5). We tested all combinations of resistant and susceptible substitutions at positions 394 (helix αE), 489 and 492 (helix αG) in human PKR and found that six out of eight combinations of human *PKR* alleles resist K3L. The two exceptions are F-Y-A (described above) and F-F-A (Fig. 3a, row 4), which is only slightly more resistant than F-Y-A to K3L, revealing a weak effect associated with the positively selected residue at position 489. Although the human and gibbon PKR backbones bear similar outcomes at all positions (Fig. 3b), the 'susceptible' human alleles still seem more resistant than the 'susceptible' gibbon alleles to vaccinia K3L, hinting at an additional K3L resistance determinant in the human PKR sequence (data not shown).

One of the most notable findings from testing the 'susceptible' and 'resistant' PKR variants was that helices αE and αG had distinct means of defeating K3L. Leu 394 resisted K3L regardless of whether human, gibbon or orangutan PKR had a 'susceptible' αG helix (Fig. 2c and Supplementary Fig. 5). The mutational profile of the αE and αG helices is therefore not strictly independent, because helix αE masks αG in terms of K3L resistance. Only when helix αE is 'susceptible' does the configuration of αG matter. Residue 394 of helix αE toggles exclusively between leucine and phenylalanine at a much lower rate than residues of helix αG, not only in primates but also among mammals in general (Supplementary Fig. 6). Our finding that Leu 394 confers overriding resistance to vaccinia K3L strongly suggests that toggling can unmask potentially adaptive substitutions in the rapidly evolving αG helix. The fact that Phe 394 is fixed in numerous species, including the New World monkeys we sampled, suggests that phenylalanine rather than leucine might confer resistance against substrate mimics different from the two we tested in this study. Therefore, toggling at position 394 reveals how a single substitution, in combination with positive selection in helix αG, might effectively increase the adaptive space that PKR can explore, greatly increasing the odds of defeating substrate mimics.

Positive selection seems to be a major evolutionary driver of many host–pathogen interactions[11,18,26]. Strong positive selection seen in both primate PKR and poxvirus K3L, and the presence of substrate

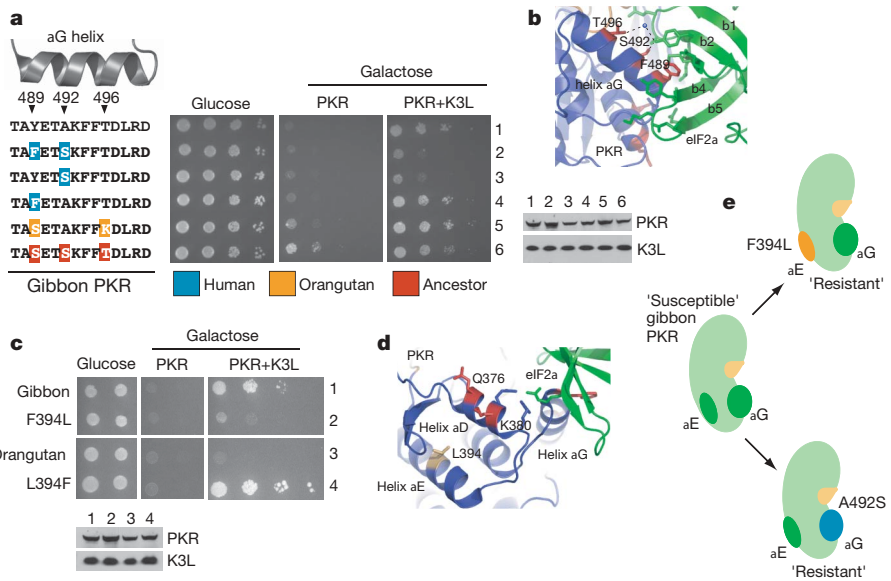


Figure 2 | Distinct surfaces of the PKR kinase domain are crucial to K3L resistance. **a**, Plasmids encoding gibbon *PKR* alleles with substitutions in the αG helix were introduced into yeast strains HM3 (eIF2α alone) and HM1 (eIF2α and K3L). Tenfold serial dilutions of transformants are shown. A corresponding immunoblot analysis is also shown with antibodies against PKR (top) and K3L (bottom). **b**, A ribbon representation of the PKR–eIF2α complex, highlighting the association of side chains of residues under positive selection with side chains of eIF2α. Phe 489, Ser 492 and Thr 496 form a face of the αG helix directly interacting with eIF2α (ref. 15). **c**, Plasmids encoding gibbon and orangutan *PKR* alleles with substitutions in the αE helix were introduced into yeast strains HM3 and HM1. Tenfold serial dilutions of transformants are shown, along with a corresponding immunoblot analysis. **d**, Residues under positive selection (Gln 376 and Lys 380) and residue Leu 394 from a ribbon representation of human PKR and eIF2α are shown[15]. **e**, Diagram showing that single substitutions in either the αE or αG helices can confer resistance against vaccinia K3L to gibbon PKR.

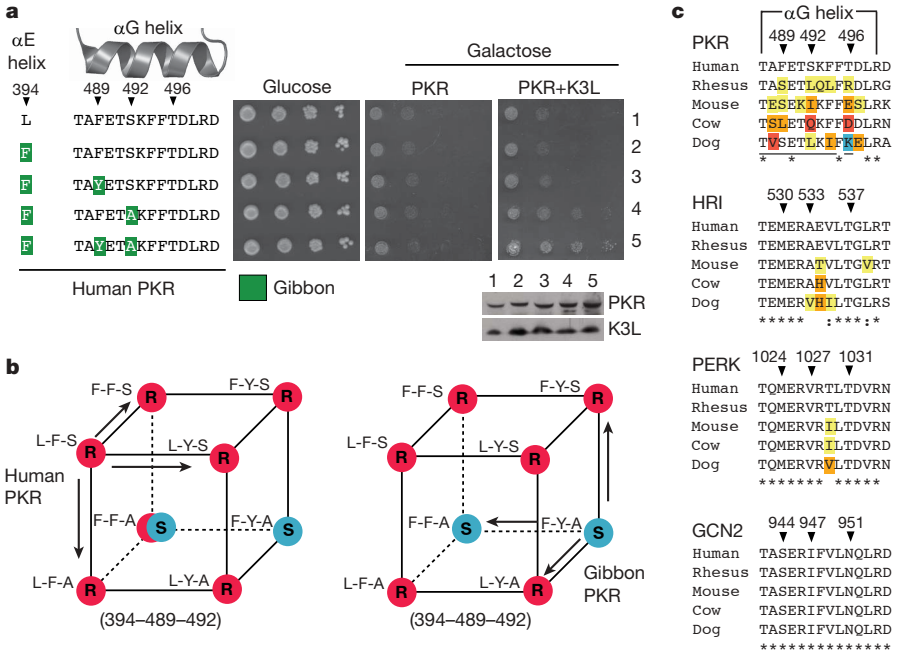


Figure 3 | PKR chimaeras reveal masking of K3L sensitivity by Leu 394. **a**, Tenfold serial dilutions of transformants expressing alleles of human *PKR* with combinations of substitutions in the αE and αG helices are shown, along with a corresponding immunoblot analysis. **b**, Phenotype 'cubes' summarizing the K3L susceptibility of alleles with all combinations of substitutions between human and gibbon PKR at positions 394, 489 and 492 from Figs 2a and 3a and Supplementary Fig. 5. Red and blue dots indicate resistance and sensitivity to K3L, respectively. With the exception of F-F-A, which shows some measure of resistance to K3L in the human background (indicated by the red crescent), each set of substitutions has similar phenotypes in the human and gibbon backgrounds. Each single substitution in wild-type human PKR results in a variant still resistant to K3L, whereas in two of three cases gibbon PKR becomes resistant (indicated by arrows). **c**, Sequence alignments of the αG helix for each member of the eIF2α kinase family (PKR, haem-regulated inhibitor (HRI), PKR-like ER kinase (PERK) and GCN2) from several mammals highlights the conservation of this region compared with the rapid evolution of PKR (black arrowheads indicate residues of the αG helix under positive selection in PKR). The frequency of substitutions in the panel at each position is indicated by a colour code (yellow for a single substitution, orange for a second, red for a third, and blue for a fourth), with the human sequence as a reference. Residues making contacts with eIF2α are indicated with lines below the PKR alignment.

mimics in unrelated viruses[27], clearly indicates that both host and viral genomes have been under intense pressure to gain advantages in these ancient and continuing evolutionary battles. The positive selection we observed in primate PKR is likely to reflect selection driven by ancient viruses with K3L-like factors that strongly influenced susceptibility to present-day mimics. For example, positive selection in the gibbon lineage driven by ancient mimics may have left gibbon PKR susceptible to vaccinia K3L. Similar trade-offs have been observed for variants of antiviral proteins under strong positive selection that might have defeated ancient retroviruses but are currently susceptible to HIV-1 (ref. 19).

Mimicry adds a layer of complexity to host–pathogen interfaces. Because PKR must distinguish an essentially unchanging substrate from rapidly evolving mimics such as K3L, it is surprising that most present-day hominoid species are resistant to vaccinia K3L (Fig. 1c). Our studies reveal evolutionary mechanisms that might allow host genes such as PKR to stave off mimicry. This strategy involves not only positive selection but also multiple discrimination interfaces (αE and αG helices) and a combinatorial outcome of resistance or susceptibility based on these surfaces, which together can increase discrimination against rapidly evolving mimics.

PKR seems well suited for molecular arms races against mimics because of a striking level of evolutionary flexibility. Because the biochemical activity of PKR depends on recognition of an unchanging substrate, strong purifying selection at the interaction interface would be expected. Indeed, other members of the eIF2α kinase family, which do not primarily serve antiviral functions and are not known to encounter viral mimicry directly, have highly conserved αG helices (Fig. 3c) and evolve under purifying selection ($dN < dS$; Supplementary Fig. 1). Despite extensive amino-acid diversity in helix αG, variants of PKR retain the ability to recognize eIF2α. The contrasting evolutionary trajectories of helix αG in the family of eIF2α kinases suggests that host factors challenged by mimics, such as PKR, rely on a high degree of flexibility to escape mimicry. We speculate that substantial selective pressures for distinguishing substrate mimics may even result in substitutions causing a decrease in substrate recognition until potential compensatory mutations might arise. Consistent with this situation was our finding that introducing an ancestral helix αG or one from orangutan into PKR from gibbon resulted in slightly compromised substrate recognition (Fig. 2a, middle, rows 5 and 6; also see Supplementary Fig. 7, middle), yet full substrate recognition was restored for helix αG from orangutan in the context of the whole protein (Fig. 1c, middle, orangutan). Compromising one function to explore a greater adaptive landscape for another function is probably a theme for genetic gains of functional novelty[28,29]. Because contending with viral mimicry can be essential for combating infectious disease, compromises to components of key cellular processes targeted by mimics[1–5,30] might be a 'hidden' evolutionary cost of such high-stakes genetic conflicts.

METHODS SUMMARY

Primate *PKR* cDNA was amplified, cloned, and sequenced from a panel of hominoids, as well as from Old World and New World monkeys. Poxvirus *K3L* sequences were retrieved from the Poxvirus Bioinformatics Resource Center (http://www.poxvirus.org). DNA sequences from each panel were aligned and used for phylogenetic and evolutionary analysis with the PAML[13] and HyPhy software packages. Structure observations of PKR were made with data coordinates from the Protein Databank (http://www.pdb.org; IDs 2A1A and 2A19) and MacPyMol software.

Variants of *PKR* and *K3L* were cloned into yeast vectors, and gene expression was driven in transformed yeast strains under a galactose-induced promoter. Yeast growth was monitored in serial-dilution series of transformants on plates containing selective medium and galactose as a carbon source. Western blots with

antibodies raised against PKR, K3L and the haemagglutinin (HA) epitope were performed to determine protein levels for yeast strains used in growth assays.

For virus infection assays, human, orangutan and white-cheeked gibbon fibroblasts were infected with 0.001 plaque-forming units per cell of wild-type or ΔK3L vaccinia virus (Copenhagen strain) for 1 h. Virus production was assessed 72 h after infection by titring cell lysates.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 19 August; accepted 8 October 2008.
Published online 30 November 2008.

1. Murphy, P. M. Molecular mimicry and the generation of host defense protein diversity. *Cell* **72,** 823–826 (1993).
2. Angot, A., Vergunst, A., Genin, S. & Peeters, N. Exploitation of eukaryotic ubiquitin signaling pathways by effectors translocated by bacterial type III and type IV secretion systems. *PLoS Pathog.* **3,** e3 (2007).
3. Benedict, C. A., Norris, P. S. & Ware, C. F. To kill or be killed: viral evasion of apoptosis. *Nature Immunol.* **3,** 1013–1018 (2002).
4. Izard, T., Tran Van Nhieu, G. & Bois, P. R. *Shigella* applies molecular mimicry to subvert vinculin and invade host cells. *J. Cell Biol.* **175,** 465–475 (2006).
5. Stebbins, C. E. & Galan, J. E. Structural mimicry in bacterial virulence. *Nature* **412,** 701–705 (2001).
6. Alcami, A. Viral mimicry of cytokines, chemokines and their receptors. *Nature Rev. Immunol.* **3,** 36–50 (2003).
7. Finlay, B. B. & McFadden, G. Anti-immunology: evasion of the host immune system by bacterial and viral pathogens. *Cell* **124,** 767–782 (2006).
8. Meurs, E. *et al.* Molecular cloning and characterization of the human double-stranded RNA-activated protein kinase induced by interferon. *Cell* **62,** 379–390 (1990).
9. Dever, T. E., Dar, A. C. & Sicheri, F. in *Translational Control in Biology and Medicine* (eds Mathews, M. B., Sonenberg, N. & Hershey, J. W. B.) 319–344 (Cold Spring Harbor Laboratory Press, 2007).
10. Langland, J. O., Cameron, J. M., Heck, M. C., Jancovich, J. K. & Jacobs, B. L. Inhibition of PKR by RNA and DNA viruses. *Virus Res.* **119,** 100–110 (2006).
11. Sawyer, S. L., Emerman, M. & Malik, H. S. Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G. *PLoS Biol.* **2,** E275 (2004).
12. Nielsen, R. *et al.* A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* **3,** e170 (2005).
13. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24,** 1586–1591 (2007).
14. Scheffler, K., Martin, D. P. & Seoighe, C. Robust inference of positive selection from recombining coding sequences. *Bioinformatics* **22,** 2493–2499 (2006).
15. Dar, A. C., Dever, T. E. & Sicheri, F. Higher-order substrate recognition of eIF2α by the RNA-dependent protein kinase PKR. *Cell* **122,** 887–900 (2005).
16. Li, Y. *et al.* On the origin of smallpox: correlating variola phylogenics with historical smallpox records. *Proc. Natl Acad. Sci. USA* **104,** 15787–15792 (2007).
17. Babkin, I. V. & Shchelkunov, S. N. The time scale in poxvirus evolution. [In Russian.] *Mol. Biol. (Mosk.)* **40,** 20–24 (2006).
18. Sawyer, S. L., Wu, L. I., Emerman, M. & Malik, H. S. Positive selection of primate TRIM5α identifies a critical species-specific retroviral restriction domain. *Proc. Natl Acad. Sci. USA* **102,** 2832–2837 (2005).
19. Kaiser, S. M., Malik, H. S. & Emerman, M. Restriction of an extinct retrovirus by the human TRIM5α antiviral protein. *Science* **316,** 1756–1758 (2007).
20. Chong, K. L. *et al.* Human p68 kinase exhibits growth suppression in yeast and homology to the translational regulator GCN2. *EMBO J.* **11,** 1553–1562 (1992).
21. Dey, M. *et al.* Mechanistic link between PKR dimerization, autophosphorylation, and eIF2α substrate recognition. *Cell* **122,** 901–913 (2005).
22. Dever, T. E. *et al.* Mammalian eukaryotic initiation factor 2α kinases functionally substitute for GCN2 protein kinase in the GCN4 translational control mechanism of yeast. *Proc. Natl Acad. Sci. USA* **90,** 4616–4620 (1993).
23. Kawagishi-Kobayashi, M., Silverman, J. B., Ung, T. L. & Dever, T. E. Regulation of the protein kinase PKR by the vaccinia virus pseudosubstrate inhibitor K3L is dependent on residues conserved between the K3L protein and the PKR substrate eIF2alpha. *Mol. Cell. Biol.* **17,** 4146–4158 (1997).
24. Langland, J. O. & Jacobs, B. L. The role of the PKR-inhibitory genes, E3L and K3L, in determining vaccinia virus host range. *Virology* **299,** 133–141 (2002).
25. Poon, A. F., Lewis, F. I., Pond, S. L. & Frost, S. D. An evolutionary-network model reveals stratified interactions in the V3 loop of the HIV-1 envelope. *PLOS Comput. Biol.* **3,** e231 (2007).
26. Kerns, J. A., Emerman, M. & Malik, H. S. Positive selection and increased antiviral activity associated with the PARP-containing isoform of human zinc-finger antiviral protein. *PLoS Genet.* **4,** e21 (2008).
27. Essbauer, S., Bremont, M. & Ahne, W. Comparison of the eIF-2α homologous proteins of seven ranaviruses (Iridoviridae). *Virus Genes* **23,** 347–359 (2001).
28. Ortlund, E. A., Bridgham, J. T., Redinbo, M. R. & Thornton, J. W. Crystal structure of an ancient protein: evolution by conformational epistasis. *Science* **317,** 1544–1548 (2007).
29. Dean, A. M. & Thornton, J. W. Mechanistic approaches to the study of evolution: the functional synthesis. *Nature Rev. Genet.* **8,** 675–688 (2007).
30. Sawyer, S. L. & Malik, H. S. Positive selection of yeast nonhomologous end-joining genes and a retrotransposon conflict hypothesis. *Proc. Natl Acad. Sci. USA* **103,** 17614–17619 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank T. Dever for yeast strains and advice; J. Tartaglia and B. Jacobs for valuable reagents; and S. Biggins and S. Furuyama for yeast expression plasmids and advice, and M. Emerman, S. Henikoff, S. Biggins, A. Turkewitz, D. Gottschling, D. Koshland, E. Smith, J. Kerns, S. Sawyer and D. Vermaak for comments and suggestions. We are supported by NIH grant AI026672 (A.P.G.) and a Searle Scholar and Burroughs Wellcome Investigator Award (H.S.M.). N.C.E. is an Ellison Medical Foundation Fellow of the Life Sciences Research Foundation.

Author Contributions N.C.E. and H.S.M. designed the study. N.C.E. performed the evolutionary analysis and yeast growth assays. S.J.C. and A.P.G. designed and performed the vaccinia infection experiments. N.C.E. and H.S.M. wrote the paper. All authors discussed and edited the manuscript.

Author Information Sequences of PKR have been deposited in Genbank under accession numbers EU733254–EU733271 and FJ374685. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to H.S.M. (hsmalik@fhcrc.org).

METHODS

Primate genomic sources. Total RNA was obtained (RNeasy kit; Qiagen) from the following cell lines from the Coriell Cell Repositories except where otherwise noted: *Homo sapiens* (human; primary human foreskin fibroblasts), *Pan troglodytes* (chimpanzee; AG06939), *Pan paniscus* (bonobo; AG05253), *Gorilla gorilla* (western lowland gorilla; AG05251), *Pongo pygmaeus pygmaeus* (Bornean orangutan; AG05252), *Hylobates agilis albibarbis* (agile gibbon; PR00773), *Hylobates leucogenys* (white-cheeked gibbon; PR01037), *Hylobates syndactylus* (island siamang; PR00722), *Cercocebus atys* (sooty mangabey; gift from C. Apetrei), *Macaca mulatta* (rhesus monkey; TF-1 cells), *Miopithecus talapoin talapoin* (talapoin monkey; PR00716), *Erythrocebus patas* (patas monkey; AG06116), *Cercopithecus aethiops* (African green monkey; COS cells), *Trachypithecus francoisi* (Francois' leaf monkey; PR01099), *Colobus guereza* (colobus monkey; PR00980), *Lagothrix lagotricha* (common woolly monkey; AG05356), *Ateles geoffroyi* (black-handed spider monkey; AG05352), *Callicebus moloch* (dusky titi; AG06115), *Aotus trivirgatus* (owl monkey; CRL-1556; American Type Culture Collection) and *Saguinus labiatus* (red-chested mustached tamarin; AG05308).

cDNA cloning and sequences. RNA (50 ng) from each primate was used for RT–PCR (SuperScript III; Invitrogen) with primers listed in Supplementary Table 10. PCR products were TA-cloned into pCR2.1 (Invitrogen) and sequenced from three different clones. The human PKR variant we cloned was identical in sequence to the GenBank entry for this gene (NM002759). The *PKR* cDNA sequence from *Callithrix jacchus* (common marmoset) was obtained by means of Blat searches of the UCSC Genome browser with PKR sequences from other New World monkeys to aid in identifying exon/intron boundaries. Other mammalian sequences of PKR, HRI, PERK and GCN2 were obtained from GenBank or by means of Blat searches of the UCSC Genome browser (http://genome.ucsc.edu). Vaccinia (Copenhagen) and variola (major) and other poxvirus K3L sequences were obtained from the Poxvirus Bioinformatics Resource Center (http://www.poxvirus.org).

PKR variants were ligated into 2-μm (pSB819; URA) and CEN (pSB146; URA) yeast *pGAL* expression plasmids by means of *Xho*I and *Not*I sites introduced into PKR primers. K3L or amino-terminal HA-epitope-tagged K3L from vaccinia virus (Copenhagen strain) was amplified by PCR, TA cloned, sequenced for accuracy, and ligated by means of *Xho*I and *Not*I sites into an integrating (pSB305; LEU) yeast expression plasmid into which a galactose promoter was also introduced. Variola major K3L sequence was synthesized by standard methods (CelTek Genes) and subcloned in the same manner as K3L from vaccinia. For comparisons of helix αE the following mammalian sequences of PKR were obtained from GenBank: *Mus musculus* (mouse; NP_035293), *Rattus norvegicus* (rat, NP_062208), *Oryctolagus cuniculus* (rabbit, NP_001075682), *Canis lupus familiaris* (dog, NP_001041600), *Equus caballus* (horse, XP_001917876), *Bos taurus* (cow, NP_835210) and *Sus scrofa* (pig, NP_999484).

Evolutionary analysis and structure observations. DNA sequences were aligned by using ClustalW with small indels trimmed on the basis of amino-acid comparisons (Supplementary Data). The generally accepted primate phylogeny (Fig. 1a) was used for evolutionary analysis, although a neighbour-joining tree generated from the alignment of PKR placed gorilla and owl monkey at different nodes (Supplementary Fig. 8). Parallel analysis with the PKR tree did not alter any results significantly (data not shown). Pairwise d*N*/d*S* analysis of eIF2α kinases, eIF2α and K3L were performed with *K*-estimator software[31]. Maximum-likelihood analysis of the larger PKR data set was performed with codeml of the PAML software package[13]. A free-ratio model allowing d*N*/d*S* variation along different branches of the phylogeny was employed to calculate d*N*/d*S* values between lineages. Two-ratio tests were performed with likelihood models comparing all branches fixed at $dN/dS = 1$ or an average d*N*/d*S* value from the whole tree applied to each branch to varying d*N*/d*S* values according to branch. Complementary analysis grouping lineages according to d*N*/d*S* values with multi-model inference (HyPhy software)[32] was also applied to the data set. We uncovered support for one recombination breakpoint in the data set by using the GARD program (HyPhy software; see Supplementary Table 5).

To detect selection in PKR, the multiple alignment was fitted to either F3x4 or F61 codon frequency models. Likelihood ratio tests (LRTs) were performed by comparing the following site-specific models (*NS* sites): M1 (neutral) with M2 (selection), M7 (neutral, β distribution of $dN/dS < 1$) with M8 (selection, beta distribution, $dN/dS > 1$ allowed), and M8a (neutral, with class of sites at $dN/dS = 1$) with M8. Similar LRTs that also account for synonymous rate variation and recombination (PARRIS; HyPhy software) were performed. Co-evolution analysis between PKR residues was also performed (Spidermonkey/BGM; HyPhy software).

PAML analysis identified sets of amino acids with high posterior probabilities (more than 0.90) for positive selection by a Bayesian approach. Similar analysis identified amino acids under positive selection with the LRTs implemented in the SLAC, FEL and REL programs (HyPhy software). Amino acids under positive selection in the kinase domain were examined using the PKR–eIF2α structure data coordinates available in the Protein Databank (PDB IDs 2A1A and 2A19; http://www.pdb.org)[15] and MacPyMol software[33].

Yeast strains and growth assays. Standard techniques were used for culturing and transforming yeast strains[34]. Strain H2557 was provided by T. Dever[21] and modified by integrating K3L or HA–K3L under the *gal* promoter at the *leu2* locus. Integration of *K3L* alleles in the resulting strains HM1 and HM2 were confirmed by PCR with primers flanking the *leu2* locus. HM3 was generated from H2557 by transforming empty pSB305 linearized with *Eco*RV. Genotypes of these strains are shown in Supplementary Table 11.

Strains HM1, HM2, HM3, HM4 and J223 (S51A allele of eIF2α; provided by T. Dever)[21] were transformed with PKR variants in 2-μm plasmid pSB819 for growth assays. Transformants were grown in YC-leu-ura medium (yeast complete minimal medium with amino acids) containing 2% glucose, then washed and plated in dilution series of $D_{600} = 3.0, 0.3, 0.03, 0.003$ with the use of a bacterial replicator (Aladin Enterprises) on YC-leu-ura medium containing either 2% glucose or 2% galactose and grown for 6 days (the human chimaera set shown in Fig. 3 and Supplementary Fig. 5 was grown for 10 days). Growth assays with PKR variants expressed from CEN plasmid pSB146 yielded consistent results (data not shown).

Western blotting. Transformants were grown to saturation in YC-leu-ura medium with 2% glucose, then washed and diluted 1:50 in YC-leu-ura medium with 2% galactose and grown for 15 h. Whole-cell lysates were prepared[35] and resolved by SDS–PAGE (12% Tris-glycine gel; Invitrogen). Proteins were transferred to nitrocellulose membranes and detected with anti-PKR antibody B-10 (1:1,000 dilution; Santa Cruz Biotechnology), anti-HA.11 (1:1,000; Covance) or a monoclonal antibody against K3L (1:2,000; a gift from J. Tartaglia).

Primate PKR infection assays. Human, orangutan and white-cheeked gibbon fibroblasts were maintained in Eagle's minimal essential medium with Earle's salts and non-essential amino acids, supplemented with 10% fetal bovine serum (Gibco), penicillin–streptomycin (100 U ml^{-1}) and 2 mM L-glutamine. Vaccinia virus Copenhagen strain (VC2)[36] and VVΔK3L (ref. 37), both obtained from B. Jacobs, were propagated and titred in BSC_{40} cells. Growth and titration of VV stocks were performed essentially as described[38], except that virus stocks were partly purified after cell lysis by centrifugation through a 36% sucrose cushion before resuspension in 1 mM Tris-HCl pH 9.0, division into aliquot, and storage at −70 °C.

Human, orangutan and gibbon fibroblasts in triplicate wells were infected with VC2 or VVΔK3L at 0.001 plaque-forming units per cell for 1 h, washed twice, and re-fed with medium. At 72 h after infection the infected cells were collected and freeze-thawed three times, and the resulting lysates were titred on BSC_{40} cells[39].

31. Comeron, J. M. K-Estimator: calculation of the number of nucleotide substitutions per site and the confidence intervals. *Bioinformatics* **15,** 763–764 (1999).
32. Pond, S. L. & Frost, S. D. A genetic algorithm approach to detecting lineage-specific variation in selection pressure. *Mol. Biol. Evol.* **22,** 478–485 (2005).
33. DeLano, W. L. *The PyMOL User's Manual* (DeLano Scientific, 2004).
34. Gietz, R. D. & Woods, R. A. Transformation of yeast by lithium acetate/single-stranded carrier DNA/polyethylene glycol method. *Methods Enzymol.* **350,** 87–96 (2002).
35. Kushnirov, V. V. Rapid and reliable protein extraction from yeast. *Yeast* **16,** 857–860 (2000).
36. Tartaglia, J. *et al.* Highly attenuated poxvirus vectors. *AIDS Res. Hum. Retroviruses* **8,** 1445–1447 (1992).
37. Beattie, E., Tartaglia, J. & Paoletti, E. Vaccinia virus-encoded eIF-2α homolog abrogates the antiviral effect of interferon. *Virology* **183,** 419–422 (1991).
38. Earl, P. L., Cooper, N., Wyatt, L. S., Moss, B. & Carroll, M. W. Preparation of cell cultures and vaccinia virus stocks. *Curr. Protocols Protein Sci.* 5, Unit 5.12 doi:10.1002/0471140864.ps0512s13 (2001).
39. Dar, A. C. & Sicheri, F. X-ray crystal structure and functional analysis of vaccinia virus K3L reveals molecular determinants for PKR subversion and substrate recognition. *Mol. Cell* **10,** 295–305 (2002).

LETTERS

Endochondral ossification is required for haematopoietic stem-cell niche formation

Charles K. F. Chan[1]*, Ching-Cheng Chen[1]*, Cynthia A. Luppen[1]*, Jae-Beom Kim[2]†, Anthony T. DeBoer[1], Kevin Wei[3], Jill A. Helms[2], Calvin J. Kuo[3], Daniel L. Kraft[1] & Irving L. Weissman[1]

Little is known about the formation of niches, local micro-environments required for stem-cell maintenance. Here we develop an *in vivo* assay for adult haematopoietic stem-cell (HSC) niche formation[1,2]. With this assay, we identified a population of progenitor cells with surface markers $CD45^{-}Tie2^{-}\alpha_V{}^{+}CD105^{+}Thy1.1^{-}$ ($CD105^{+}Thy1^{-}$) that, when sorted from 15.5 days post-coitum fetal bones and transplanted under the adult mouse kidney capsule, could recruit host-derived blood vessels, produce donor-derived ectopic bones through a cartilage intermediate and generate a marrow cavity populated by host-derived long-term reconstituting HSC (LT-HSC). In contrast, $CD45^{-}Tie2^{-}\alpha_V{}^{+}CD105^{+}Thy1^{+}$ ($CD105^{+}Thy1^{+}$) fetal bone progenitors form bone that does not contain a marrow cavity. Suppressing expression of factors involved in endochondral ossification, such as osterix and vascular endothelial growth factor (VEGF), inhibited niche generation. $CD105^{+}Thy1^{-}$ progenitor populations derived from regions of the fetal mandible or calvaria that do not undergo endochondral ossification formed only bone without marrow in our assay. Collectively, our data implicate endochondral ossification, bone formation that proceeds through a cartilage intermediate, as a requirement for adult HSC niche formation.

Growth and renewal in many tissues are initiated by stem cells, supported by the niches in which they reside[1–3]. Although recent work has begun to describe functional interactions between stem cells and their niches, little is known about the formation of stem-cell niches. Identification of the cells and processes that can generate, sustain and influence the HSC niche and haematopoiesis are critical for our understanding of normal haematopoiesis; stem-cell homing, trafficking and differentiation; and haematopoietic pathology[4–12]. There is a need for modular systems in which the cellular and molecular components of a niche can be genetically modified and studied *in vivo*.

We have established an *in vivo* assay that allows functional assessment of the formation and maintenance of HSC niches at an ectopic site. We hypothesized that circulating HSC would colonize a non-haematopoietic location and establish functional haematopoiesis if appropriate niche components were present. We selected the subcapsular site of the kidney because it possesses a rich vascular supply, supports several kinds of tissue engraftment and is not known to contain HSCs. In mice, HSCs are not detectable in the limb bone rudiment until 17.5 days post-coitum (d.p.c.)[13]. In our initial experiments, we showed that transplantation of 14.5 d.p.c. fetal bones under the kidney capsule, into either green fluorescent protein (GFP) transgenic or CD45 congenic hosts, resulted in the formation of donor-derived bones with host-derived marrow and HSCs (Supplementary Fig. 2). This result indicates that 14.5 d.p.c. fetal bones contain elements that can initiate an ectopic niche. To determine if the fetal bones must be intact for niche initiation, we dissociated the 14.5 d.p.c. fetal bones into a single cell suspension, embedded the suspension in Matrigel, and introduced it under the kidney capsule. The suspension generated both cartilaginous and membranous bones that were populated with phenotypic and functional HSC (Fig. 1a–e and

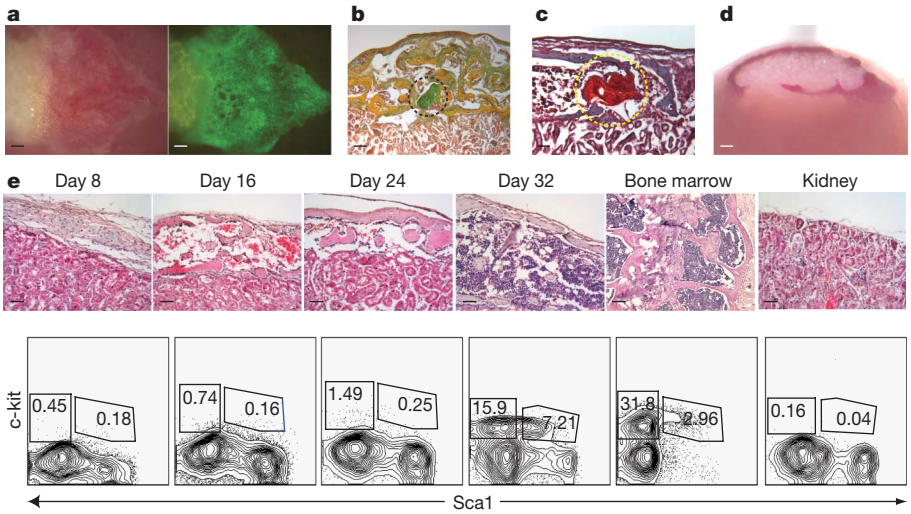


Figure 1 | Fetal bone cells can initiate an ectopic HSC niche. a, Ectopic bone formed by GFP-labelled, 14.5 d.p.c. fetal bone cells 32 days after subrenal capsule transplant (scale bar, 500 μm). **b**, Representative section of graft stained with pentachrome (yellow, osteoid; greenish blue, cartilage), cartilaginous region in black circle (scale bar, 100 μm). **c**, Safranin-O stain of adjacent section, red-staining cartilage matrix in yellow circle (scale bar, 200 μm). **d**, Alizarin Red stain for calcified tissue (scale bar, 500 μm). **e**, Time course study of haematopoietic components during ectopic niche formation. Haematoxylin and eosin staining (upper panel; scale bar, 100 μm), representative FACS profiles of LT-HSC ($CD45^{+}lineage^{-}c\text{-}kit^{+}Sca1^{+}CD150^{+}$) frequency that were pregated for live, $CD45^{+}lineage^{-}$ cells (lower panel). Days after transplantation are indicated; $n = 4$ for each time point.

[1]Department of Pathology, Developmental Biology and Institute for Stem Cell Biology and Regenerative Medicine, Stanford University, California, USA. [2]Department of Surgery, Division of Plastic and Reconstructive Surgery, Stanford University, California, USA. [3]Department of Hematology, Stanford University, California, USA. †Present address: Caliper Life Sciences, 2061 Challenger Drive, Alameda, California 94501, USA.
*These authors contributed equally to this work.

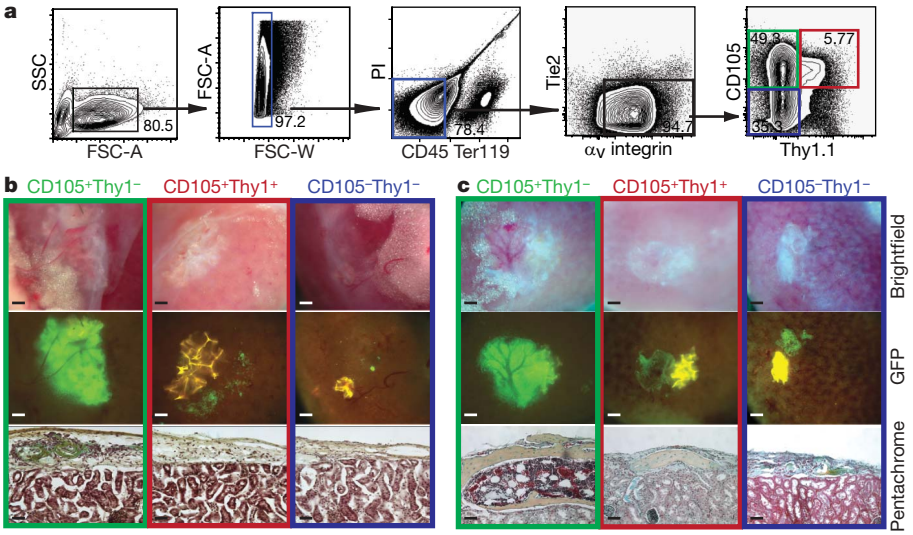


Figure 2 | CD105$^+$Thy1$^-$ population forms ectopic HSC niche through a cartilaginous intermediate. **a**, Representative FACS profile of homogenized 15.5 d.p.c. fetal bones pre-gated for live CD45$^-$Ter119$^-$ cells showing CD105$^+$Thy1$^-$, CD105$^+$Thy1$^+$ and CD105$^-$Thy1$^-$ populations (green, red and blue gates, respectively). **b**, **c**, Two thousand double-sorted fetal bone cells from each fraction were injected under the renal capsule and harvested 16 days ($n = 4$) (**b**) or 32 days ($n = 13$) (**c**) after transplantation. Bright-field (upper panel) and GFP images (middle panel) of explanted ectopic grafts. Pentachrome staining of transverse sections through grafts at 16 and 32 days (lower panel). (Scale bar in upper and middle panels, 500 μm; in lower panel, 100 μm.)

Supplementary Figs 1a, b and 5). To distinguish donor from host tissue, we transplanted either fetal bones from GFP transgenic mice or wild-type fetal bone suspension transduced with lentiviral-GFP. The haematopoietic and vascular components within the ectopic niche were host derived; non-haematopoietic and non-vascular components, including bone and cartilage, were donor derived (Supplementary Figs 2–4). The engraftment and activity of host HSC within these ectopic niches has been verified by surface marker phenotype and functional long-term engraftment assays in secondary recipients (Supplementary Figs 1a, b and 5). We did not detect HSCs either in the ungrafted kidneys of transplanted mice or in kidneys transplanted with Matrigel only (Supplementary Fig. 1a, b). To determine the kinetics of HSC colonization relative to ectopic bone formation, we evaluated both the presence of LT-HSC and histological parameters of bone formation at 8-day intervals over a period of 32 days. Donor-derived bone was present at day 16 post-transplant, coincident with the appearance of erythrocytes (Fig. 1e) and host-derived PECAM$^+$ vasculature (Supplementary Figs 3 and 4). By day 24, c-kit$^+$ progenitors appeared; however, host-derived HSC were not detected until day 32. The day 32 grafts were structurally similar to normal bones with regions of cartilaginous, compact and trabecular bone (Fig. 1b–e). The presence of HSCs was found to be stable after the ectopic niche was established (data not shown).

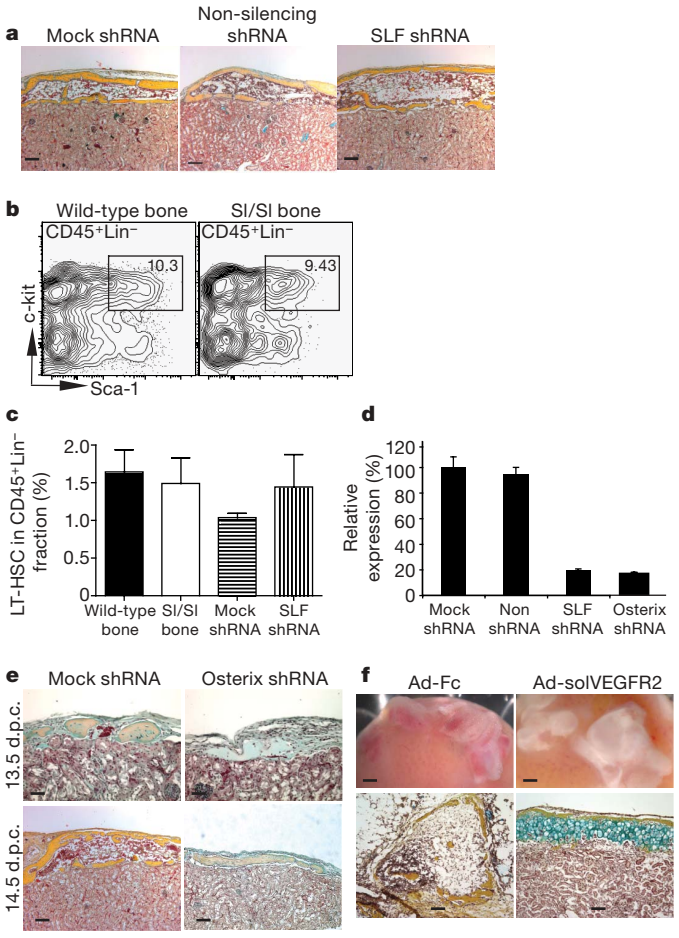


Figure 3 | Niche formation is dependent on endochondral ossification. **a**–**c**, Suppression of SLF expression in fetal bone cells did not alter osteogenesis or niche formation. Fetal bone cells (14.5 d.p.c.) were transduced with lentivirus as indicated, and 2,000 sorted GFP$^+$ cells were injected under the renal capsule. **a**, Paraffin-embedded sections were obtained 32 days after transplantation and stained with pentachrome. **b**, Representative FACS profiles of pre-gated, live CD45$^+$lineage$^-$ cells harvested from intact 15.5 d.p.c. normal (wild type) or mutant (Sl/Sl) fetal bones 40 days after transplantation ($n = 4$). **c**, Frequency of LT-HSC in different ectopic niches shown as mean values ± s.e.m. (wild-type bone, $n = 3$; Sl/Sl bone $n = 4$; sh-Mock, $n = 6$; sh-SLF, $n = 3$). **d**, Knockdown efficiency of sh-SLF and sh-osterix was determined by quantitative reverse transcription PCR (qRT–PCR) in a 1A5 osteoblast cell line. Relative expression percentage is shown as mean values ± s.d. ($n = 3$). **e**, Suppression of osterix expression disrupted osteogenesis and niche formation. Fetal bone cells from 13.5 d.p.c. (upper panel) or 14.5 d.p.c. (lower panel) were transduced with the lentivirus indicated and 2,000 sorted GFP$^+$ cells were transplanted ($n = 4$). **f**, VEGF is required for formation of bone marrow cavity. Adult C57Bl/6 mice were intravenously injected with (multiplicity of infection = 10^8) adenovirus constructs the expressing either mouse Fc (Ad-Fc) or the soluble ectodomain of VEGFR2 (Ad-solVEGFR2) ($n = 3$). Two days after virus injection, 13.5 d.p.c. fetal bone elements were transplanted under the renal capsule. Paraffin-embedded sections were obtained 25 days after transplantation and stained with pentachrome. (Scale bars in brightfield images, 500 μm; in pentachrome-stained images, 100 μm.)

To characterize the progenitors responsible for bone formation and their role in haematopoietic niche formation and maintenance further, we fractionated fetal-bone cell suspensions with a panel of cell-surface markers for: putative mesenchymal stem cells (CD105 and Thy1.1); the angiopoietin receptor that is on a variety of haematopoietic and vascular cells (Tie2) (ref. 8); haematolymphoid cells (CD45); and a vascular integrin (α_V integrin)[14]. CD105 and Thy1.1 are also expressed on haematopoietic and endothelial cells, but these cell types were negatively gated from the skeletal progenitors with CD45 and Tie2, respectively. By transplanting distinct donor fractions sorted by flow cytometry, we identified the minimal progenitor population required for the formation of a functional ectopic HSC niche. We found that $CD105^+Thy1^-$ progenitors consistently gave rise to bone with incorporated HSC-containing-marrow, whereas equal numbers of $CD105^+Thy1^+$ progenitors resulted in bone formation without marrow (Fig. 2a–c and Supplementary Fig 1c). $CD105^-Thy1^-$ populations did not form bones or niches efficiently. Thus far, we have not observed marrow formation without bone, which suggests that bone forming cells such as osteoblasts, osteoblast progenitors and/or osteoblast-associated cells may be at least indirectly or structurally important for niche formation. It is important to note, however, that neither osteoblasts nor osteoblast progenitors alone were sufficient to initiate niche formation and HSC engraftment as the $CD105^+Thy1^+$ population generated marrowless bone. To differentiate between the $CD105^+Thy1^-$ bone niche-generating and the $CD105^+Thy1^+$ bone-generating populations, we conducted time-course studies to compare the respective mechanisms of bone formation. We transplanted equal numbers of each population and harvested the grafted kidneys at 16 and 32 days post-transplantation for histological characterization. We found that only $CD105^+Thy1^-$ progenitors formed bones through a cartilage intermediate, also known as endochondral ossification, whereas $CD105^+Thy1^+$ progenitors formed bones without a detectable cartilage intermediate (Fig. 2b, c). Furthermore, we found that expression of osteocalcin, a marker of mature osteoblasts[15], was fivefold higher in $CD105^+Thy1^+$ populations derived from 15.5 d.p.c. fetal bones (Supplementary Fig. 1d, e). These results suggest that $CD105^+Thy1^+$ progenitors may have lost chondrocyte potential[16].

To assess the importance of specific candidate factors for niche development and HSC maintenance, we suppressed the expression of steel factor (SLF), osterix and VEGF. SLF is essential for adult haematopoiesis and HSC activity[17], and is expressed both by mature osteoblasts[18] and endothelial cells[19]. We inhibited expression of SLF by transducing fetal bone suspensions with a GFP-labelled, SLF-specific short hairpin RNA (shRNA) lentiviral vector before renal capsule transplant (Fig. 3d). We observed normal osteogenesis and niche formation when SLF expression was inhibited (Fig. 3a, c). HSC engraftment was similar between mock-transduced and

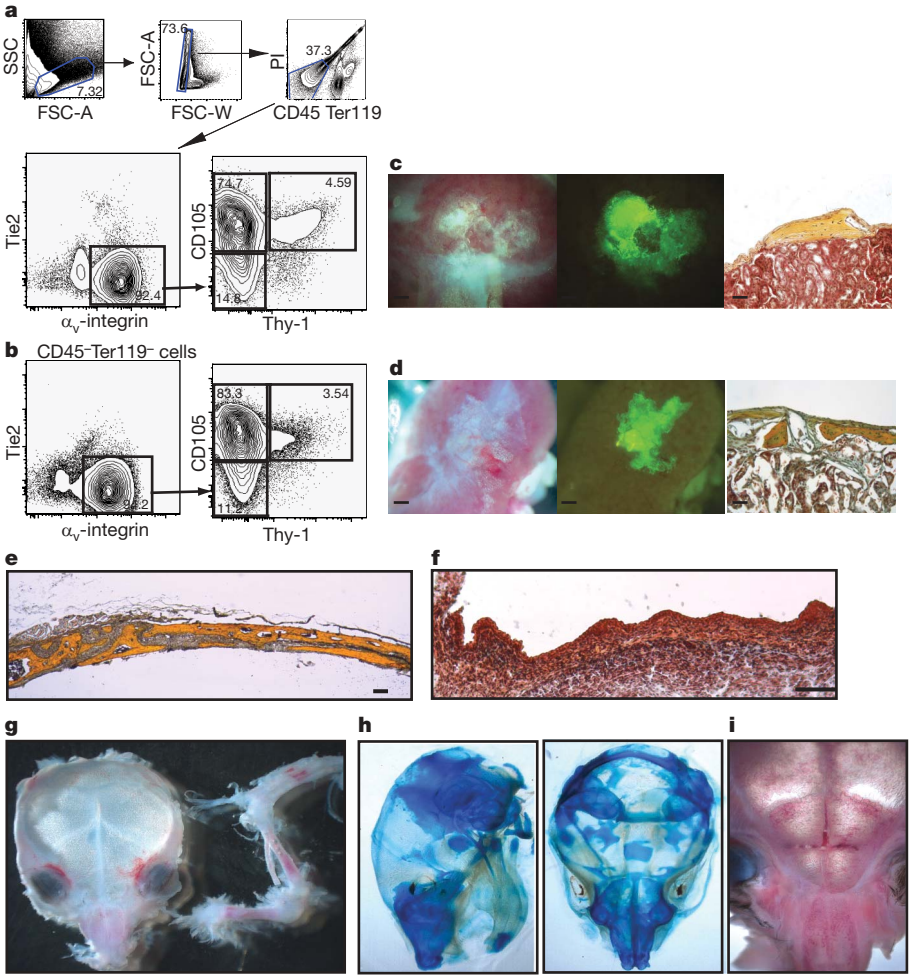


Figure 4 | Skeletal progenitors from mandible and calvaria do not form HSC niches efficiently. **a**, **b**, Representative FACS profile of homogenized 15.5 d.p.c. mandible ($n = 6$; PI, propidium iodine) (**a**) or calvaria ($n = 6$) (**b**) pre-gated on live $CD45^-Ter119^-$ cells. **c**, **d**, Two thousand sorted $GFP^+CD105^+Thy1^-$. cells from mandible (**c**) or calvaria (**d**) were transplanted under renal capsule and harvested after 32 days. **e**, Pentachrome-stained cross section of mouse parietal bone at 4 weeks. **f**, Pentachrome-stained cross section of equivalent area in 15 d.p.c. fetal calvaria. **g–i**, Marrow pockets in calvaria are concentrated in facial areas corresponding to cartilaginous regions ($n = 3$). Limb bones are juxtaposed to skull for comparison. **h**, Dorsal and lateral views of newborn calvaria show cartilaginous regions that stain with alcian blue. (Scale bar in brightfield and GFP images, 500 μm; in pentachrome images, 100 μm, in (**f**), 25 μm.)

SLF-deficient grafts, consistent with our previous observation that SLF production is not essential for haematopoiesis during the fetal period[20,21]. To confirm the results of the knockdown, we transplanted intact 14.5 d.p.c. fetal bones from SLF null mutant (Sl/Sl) and again did not observe defects in either osteogenesis or niche formation (Fig. 3b, c). This indicates that the SLF produced by skeletal progenitors and mature bone tissue is not required for initiation or maintenance of niche activity. We next silenced expression of osterix (Fig. 3e), a transcription factor necessary for endochondral ossification, in fetal bone suspensions using lentivirus vector for osterix-specific shRNA[21,22]. The osterix knockdown severely inhibited osteogenesis and abolished niche formation, underscoring the dependence of niche formation on the process of endochondral ossification. Because perichondrial cells and chondrocytes in the developing limb express high levels of VEGF, and vascular invasion is critical to endochondral ossification[23], we tested whether VEGF activity was required for niche formation. We suppressed endogenous VEGF by injecting the host mice with adenovirus expressing soluble VEGF receptor (soluble Flk1; Ad-solVEGFR2), a known VEGF-inhibiting reagent[24]. We found that endochondral ossification was disrupted when 13.5 d.p.c. fetal bones were transplanted into hosts treated with Ad-solVEGFR2 but not into hosts treated with control virus (Ad-Fc). The 13.5 d.p.c. fetal bones grafted into Ad-solVEGFR2 mice displayed an accumulation of chondrocytes without perfusing vasculature and marrow cavity, although total HSC numbers in the host were not significantly affected (Fig. 3f)[25]. To test further the hypothesis that niche formation requires endochondral ossification, we isolated $CD105^+Thy1^-$ progenitors from the regions of fetal mandible and calvaria that form bone primarily through intramembranous ossification[26], which occurs without a cartilaginous intermediate. In these experiments $CD105^+Thy1^-$ mandibular and calvarial progenitors could only form marrowless bones, even 60 days after transplantation (Fig. 4). These results suggest that endochondral ossification is necessary for niche formation.

A better functional understanding of the HSC niche was gained by observing how it is formed. In this study, $CD105^+Thy1^-$ skeletal progenitors isolated from fetal limb bones initiated ectopic HSC niche formation. The progenitors gave rise to donor-derived chondrocytes, which recruited host-derived vasculature into the centre of the developing bone graft. As endochondral ossification proceeded, the recruited vasculature facilitated the filling of the niche with host-derived haematopoietic cells: first erythroid and myeloid, then $c\text{-}kit^+$ progenitors, and finally the HSCs. We do not yet know if the $CD105^+Thy1^-$ cells represent a homogeneous population, or phenotypically similar but heterogeneous subsets. In agreement with our findings, Sacchetti *et al.* recently identified $CD146^+$ subendothelial cells residing in adult human bone marrow stroma that can generate both bone and marrow when transplanted under the skin of immunodeficient mice[27]. Although their study did not verify the presence of LT-HSC in the subdermal grafts, it is possible that osteoprogenitors in the adult marrow are involved in maintaining the niche. In addition to identifying the fetal-bone-derived skeletal progenitors that are capable of both endochondral ossification and HSC niche formation, our study provides a functional framework by which, in combination with previously described methods[6,7], HSC–niche interactions can be further investigated at the cellular level.

METHODS SUMMARY

C57/BL6 CD45.1/2 congenic mouse strains were derived and maintained in our laboratory. Timed embryos from GFP transgenic HZ mice were used in most of the fetal bone transplantation studies. Sl/+ mice were purchased from the Jackson laboratory.

Skeletal progenitors were isolated from fetal bones (humerus, radius, tibia, femur, pelvis, mandible without the condyle, and the individual frontal and parietal bones by collagenase digestion). Next, they were stained with antibodies against CD45, Tie2, α_V integrin, CD105 and Thy1.1 for fractionation by fluorescence-activated cell sorting (FACS). Sorted and unsorted skeletal progenitors were then injected underneath the renal capsule of 8- to 12-week-old anaesthetized mice.

SLF and osterix-specific shRNA knockdown constructs, active lentiviral stock and non-silencing shRNA constructs were generated as previously described[28] (Supplementary Table 1). Fetal bone cell suspensions were transduced for 48 h with specific shRNA vectors or control, sorted for GFP expression and transplanted as described.

To assess HSC engraftment in ectopic niches, grafted regions were dissected from kidney and crushed by mortar and pestle. Dissociated cells were stained with fluorochrome-conjugated antibodies against CD45, lineage (CD3, CD4, CD5, CD8, B220, Gr-1, Mac-1 and Ter119), c-kit, Sca-1 and CD150 for FACS analysis. Sorted KLS, $CD150^+$ LT-HSC were transplanted into lethally irradiated (800 rad delivered in split dose) 8- to 12-week-old congenic recipients by intravenous injection for functional analysis. Peripheral blood was obtained from the tail vein at 4 and 23 weeks after LT-HSC transplantation to assess donor-derived contributions by FACS.

Histological analyses of endochondral ossification were performed on sections that were obtained from either fresh frozen, optimal cutting temperature media (OCT)-embedded or formaldehyde-fixed, paraffin-embedded specimens. Representative sections were stained with either haematoxylin and eosin, Movat's modified pentachrome[29], Safranin-O or Alizarin Red stains, depending on the experiments.

RNA was extracted from sorted cells using Trizol (Invitrogen) or RNeasy RNA isolation kits (Qiagen) and was reverse-transcribed into complementary DNA (cDNA) with SuperscriptRT III (Invitrogen). SYBR Green Universal Master Mix and a GeneAmp 7000 or 7500 fast sequence detection system (Applied Biosystems) were used for real-time PCR with the primers listed in Supplementary Table 2. Relative expression was calculated for each gene by the comparative CT ($2^{-\Delta\Delta C_T}$) method with β-actin for normalization.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 16 June; accepted 10 October 2008.
Published online 10 December 2008.

1. Moore, K. A. & Lemischka, I. R. Stem cells and their niches. *Science* **311,** 1880–1885 (2006).
2. Adams, G. B. & Scadden, D. T. The hematopoietic stem cell in its place. *Nature Immunol.* **7,** 333–337 (2006).
3. Wilson, A. & Trumpp, A. Bone-marrow haematopoietic-stem-cell niches. *Nature Rev. Immunol.* **6,** 93–106 (2006).
4. Weissman, I. L. in *Novartis Foundation Symposium* No. 265, 35–50; discussion 50–34 92–37 (Novartis Foundation, 2005).
5. Kiel, M. J. *et al.* SLAM family receptors distinguish hematopoietic stem and progenitor cells and reveal endothelial niches for stem cells. *Cell* **121,** 1109–1121 (2005).
6. Calvi, L. M. *et al.* Osteoblastic cells regulate the haematopoietic stem cell niche. *Nature* **425,** 841–846 (2003).
7. Zhang, J. *et al.* Identification of the haematopoietic stem cell niche and control of the niche size. *Nature* **425,** 836–841 (2003).
8. Arai, F. *et al.* Tie2/angiopoietin-1 signaling regulates hematopoietic stem cell quiescence in the bone marrow niche. *Cell* **118,** 149–161 (2004).
9. Wright, D. E. *et al.* Physiological migration of hematopoietic stem and progenitor cells. *Science* **294,** 1933–1936 (2001).
10. Whitlock, C. A., Tidmarsh, G. F., Muller-Sieburg, C. & Weissman, I. L. Bone marrow stromal cell lines with lymphopoietic activity express high levels of a pre-B neoplasia-associated molecule. *Cell* **48,** 1009–1021 (1987).
11. Moore, K. A., Ema, H. & Lemischka, I. R. *In vitro* maintenance of highly purified, transplantable hematopoietic stem cells. *Blood* **89,** 4337–4347 (1997).
12. Dexter, T. M., Allen, T. D. & Lajtha, L. G. Conditions controlling the proliferation of haemopoietic stem cells *in vitro*. *J. Cell. Physiol.* **91,** 335–344 (1977).
13. Christensen, J. L., Wright, D. E., Wagers, A. J. & Weissman, I. L. Circulation and chemotaxis of fetal hematopoietic stem cells. *PLoS Biol.* **2,** E75 (2004).
14. Cheresh, D. A. Human endothelial cells synthesize and express an Arg-Gly-Asp-directed adhesion receptor involved in attachment to fibrinogen and von Willebrand factor. *Proc. Natl Acad. Sci. USA* **84,** 6471–6475 (1987).
15. Geoffroy, V., Ducy, P. & Karsenty, G. A. PEBP2 alpha/AML-1-related factor increases osteocalcin promoter activity through its binding to an osteoblast-specific cis-acting element. *J. Biol. Chem.* **270,** 30973–30979 (1995).
16. Akiyama, H. *et al.* Osteo-chondroprogenitor cells are derived from Sox9 expressing precursors. *Proc. Natl Acad. Sci. USA* **102,** 14665–14670 (2005).
17. Fleischman, R. A. & Mintz, B. Prevention of genetic anemias in mice by microinjection of normal hematopoietic stem cells into the fetal placenta. *Proc. Natl Acad. Sci. USA* **76,** 5736–5740 (1979).
18. Blair, H. C. *et al.* Parathyroid hormone-regulated production of stem cell factor in human osteoblasts and osteoblast-like cells. *Biochem. Biophys. Res. Commun.* **255,** 778–784 (1999).
19. Aye, M. T. *et al.* Expression of stem cell factor and c-kit mRNA in cultured endothelial cells, monocytes and cloned human bone marrow stromal cells (CFU-RF). *Exp. Hematol.* **20,** 523–527 (1992).

20. Ikuta, K. & Weissman, I. L. Evidence that hematopoietic stem cells express mouse c-kit but do not depend on steel factor for their generation. *Proc. Natl Acad. Sci. USA* **89,** 1502–1506 (1992).
21. Nakashima, K. *et al.* The novel zinc finger-containing transcription factor osterix is required for osteoblast differentiation and bone formation. *Cell* **108,** 17–29 (2002).
22. Koga, T. *et al.* NFAT and Osterix cooperatively regulate bone formation. *Nature Med.* **11,** 880–885 (2005).
23. Zelzer, E. *et al.* VEGFA is necessary for chondrocyte survival during bone development. *Development* **131,** 2161–2171 (2004).
24. Jacobi, J. *et al.* Adenoviral gene transfer with soluble vascular endothelial growth factor receptors impairs angiogenesis and perfusion in a murine model of hindlimb ischemia. *Circulation* **110,** 2424–2429 (2004).
25. Maes, C. *et al.* Impaired angiogenesis and endochondral bone formation in mice lacking the vascular endothelial growth factor isoforms VEGF164 and VEGF188. *Mech. Dev.* **111,** 61–73 (2002).
26. Hall, B. K. & Miyake, T. All for one and one for all: condensations and the initiation of skeletal development. *Bioessays* **22,** 138–147 (2000).
27. Sacchetti, B. *et al.* Self-renewing osteoprogenitors in bone marrow sinusoids can organize a hematopoietic microenvironment. *Cell* **131,** 324–336 (2007).
28. Metz, M. *et al.* Mast cells can enhance resistance to snake and honeybee venoms. *Science* **313,** 526–530 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank B. Péault for suggestions; L. Jerabek for laboratory management; C. Muscat for antibody production and conjugation;Y. Park, J. K. Lee and A. K. Bershad for technical support; and T. Serwold and L. Richie for advice and reading the manuscript. This study was supported in part by a USPHS National Institutes of Health (NIH) grant (2R01HL058770-08) and in part by a NIH grant (5R01CA086065-09), terminated by NIH Study Section, and in part by the Virginia and Daniel K. Ludwig Professorship to I.L.W. C.K.F.C. and C.A.L. are supported by an NIH Regenerative Medicine training grant. C.-C.C. is supported by an NIH Pathway to Independence award (5K99HL087936-02). D.L.K. is supported by an NIH Career Development award (K08-HL076335) and a Hope Street Kids research award, K. W. and C. J. K. are supported by NIH grants (1R01HL074267-01 and 1R01NS052830-01).

Author Contributions C.K.F.C. initiated the project and, with C.-C.C. and I.L.W., supervised the study, C.K.F.C., C.-C.C., C.A.L., D.L.K., J.B.K. and I.L.W. conceived and designed the experiments. C.K.F.C., C.-C.C., C.A.L., D.L.K., J.B.K., A.T.D. and K.W. performed the experiments. C.K.F.C., C.-C.C., C.A.L., D.L.K., J.B.K., A.T.D., J.A.H., C.J.K., and I.L.W. analysed the data. C.K.F.C., C.-C.C., I.L.W., C.J.K. and J.A.H. contributed reagents/materials/analysis tools. C.K.F.C., C.-C.C., C.A.L., D.L.K. and I.L.W. prepared the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to C.K.F.C. (chazchan@stanford.edu) or C.-C.C. (c3chen@stanford.edu).

METHODS

Mice. C57BL/Ka-Thy1.1-CD45.1 (HZ), C57BL/Ka-Thy1.1-CD45.1 (BA), C57BL/Ka-Thy1.2-CD45.1 (Ly5.2) and C57BL/Ka-Thy1.2-CD45.1 (B6) strains were derived and maintained in our laboratory. Timed embryos from GFP transgenic HZ mice were used in most of the fetal bone transplantation studies. shRNA studies used fetal bones from BA or HZ embryos. Eight- to twelve-week-old $RAG2^{-/-}\gamma c^{-/-}$ mice in B6 background were used as recipients for fetal bone transplantation. B6 or Ly5.2 mice were used as recipients in LT-HSC transplantation assays. Sl/+ mice were purchased from the Jackson laboratory. Sl/Sl embryos were screened by genomic PCR before transplantation. All animals were maintained in Stanford University Laboratory Animal Facility in accordance with Stanford Animal Care and Use Committee and National Institutes of Health guidelines.

Isolation and transplantation of fetal skeletal progenitors. Fetal skeletal elements (humerus, radius, tibia, femur, pelvis, mandible without the condyle, and the individual frontal and parietal bones) were dissected from euthanized mouse fetuses and digested in collagenase with DNase at 37 °C for 40 min under constant agitation. After collagenase treatment, undigested materials were gently triturated by repeated pipetting. Total dissociated cells were filtered through 40-μm nylon mesh, pelleted at 200*g* at 4 °C, resuspended in staining media (2% fetal calf serum in PBS), blocked with rat IgG and stained with fluorochrome-conjugated antibodies against CD45, Tie2, α_V integrin, CD105 and Thy1.1 for purification by flow cytometry sorting. Sorted and unsorted skeletal progenitors were pelleted and resuspended in 2 μl of Matrigel, then injected underneath the renal capsule of 8- to 12-week-old anaesthetized mice.

shRNA transduction. SLF and osterix-specific shRNA knockdown constructs and active lentiviral stocks were generated as previously described[28] (Supplementary Table 1). Fetal bone cells were resuspended in αMEM medium with 15% FCS and transduced with lentiviral vectors carrying SLF or osterix-targeting shRNA, non-silencing shRNA or empty vector. Forty-eight hours after transduction, cells were sorted for GFP expression and transplanted as described. Suppression of SLF or osterix by shRNA was verified by qRT–PCR in an osteoblastic cell line, 1A5, that expresses SLF and osterix (a gift from J.-B.K.).

Analysis of HSC engraftment in ectopic niches. Kidneys were removed from host mice, kept on ice and imaged. The grafted regions were dissected from the surface of the host kidney and then crushed between frosted, precleaned micro-slides (VWR). Homogenized tissues were filtered through 40-μm nylon mesh. Mononuclear cells were isolated using Histopaque 1107 according to the manufacturer's protocol. Cells were washed with staining media, then blocked with rat IgG and stained with fluorochrome-conjugated antibodies purchased from eBiosciences against CD45, lineage (CD3, CD4, CD5, CD8, B220, Gr-1, Mac-1 and Ter119), c-kit, Sca1 and CD150 for flow cytometry analysis.

LT-HSC functional assay. Sorted LT-HSC (lineage-, c-kit^+, Sca-1+, $CD150^+$, $CD45^+$) and 3×10^5 helper marrow cells were transplanted into lethally irradiated (800 rad delivered in split dose) 8- to 12-week-old congenic recipients by injection into the retro-orbital sinus. Peripheral blood was obtained from the tail vein at 4 and 23 weeks after LT-HSC transplantation to assess donor-derived contributions by flow cytometry.

Histological analysis of endochondral ossification. Dissected specimens were fixed in 2% PFA at 4 °C overnight, then decalcified in 0.4 M EDTA in PBS (pH 7.2) at 4 °C for 2 weeks. Specimens were then processed for embedding in paraffin (by dehydration in alcohol and xylene) or OCT (by cryoprotection in sucrose) and sectioned. Representative sections were stained with either haematoxylin and eosin, Movat's modified pentachrome[29], Safranin-O or Alizarin Red stains, depending on the experiments.

Immunofluorescent histology. Immunofluorescence on cryopreserved ectopic bone specimens was performed using an M.O.M. immunodetection kit from Vector Laboratories according to the manufacturer's instructions. Briefly, specimens were treated with a blocking reagent, than probed with monoclonal antibody at 4 °C overnight. Specimens were next washed with PBS, probed with alexa-dye-conjugated antibodies, washed, coverslipped and imaged with a Leica DMI6000B inverted microscope system. Rat anti-PECAM(CD31) monoclonal antibody was purchased from Abcam. Alexa-dye-conjugated Goat anti-Rat secondary antibodies were purchased from Molecular Probes.

RNA extraction and qRT–PCR. RNA was extracted from sorted cells using Trizol (Invitrogen) or RNeasy RNA isolation kits (Qiagen) and was reverse-transcribed into cDNA with SuperscriptRT III (Invitrogen). SYBR Green Universal Master Mix and a GeneAmp 7000 or 7500 fast sequence detection system (Applied Biosystems) were used for real-time PCR with the primers listed in Supplementary Table 2. Relative expression was calculated for each gene by the 2-ΔΔCT method with β-actin for normalization.

29. Garvey, W. *et al.* Improved Movat pentachrome stain. *Stain Technol.* **61**, 60–62 (1986).

LETTERS

Pulsed contractions of an actin–myosin network drive apical constriction

Adam C. Martin[1,2], Matthias Kaschube[3,4] & Eric F. Wieschaus[1,2]

Apical constriction facilitates epithelial sheet bending and invagination during morphogenesis[1,2]. Apical constriction is conventionally thought to be driven by the continuous purse-string-like contraction of a circumferential actin and non-muscle myosin-II (myosin) belt underlying adherens junctions[3–7]. However, it is unclear whether other force-generating mechanisms can drive this process. Here we show, with the use of real-time imaging and quantitative image analysis of *Drosophila* gastrulation, that the apical constriction of ventral furrow cells is pulsed. Repeated constrictions, which are asynchronous between neighbouring cells, are interrupted by pauses in which the constricted state of the cell apex is maintained. In contrast to the purse-string model, constriction pulses are powered by actin–myosin network contractions that occur at the medial apical cortex and pull discrete adherens junction sites inwards. The transcription factors Twist and Snail differentially regulate pulsed constriction. Expression of *snail* initiates actin–myosin network contractions, whereas expression of *twist* stabilizes the constricted state of the cell apex. Our results suggest a new model for apical constriction in which a cortical actin–myosin cytoskeleton functions as a developmentally controlled subcellular ratchet to reduce apical area incrementally.

During *Drosophila* gastrulation, apical constriction of ventral cells facilitates the formation of a ventral furrow and the subsequent internalization of the presumptive mesoderm. Although myosin is known to localize to the apical cortex of constricting ventral furrow cells[8–11], it is not known how myosin produces force to drive constriction. Understanding this mechanism requires a quantitative analysis of cell and cytoskeletal dynamics. We therefore developed methods to reveal and quantify apical cell shape with Spider–GFP, a green fluorescent protein (GFP)-tagged membrane-associated protein that outlines individual cells (Fig. 1a, b, Supplementary Fig. 1 and Supplementary Video 1)[12]. Ventral cells were constricted to about 50% of their initial apical area before the onset of invagination and continued to constrict during invagination (Fig. 1c, e). Although the average apical area steadily decreased at a rate of about 5 $\mu m^2\ min^{-1}$, individual cells showed transient pulses of rapid constriction that exceeded 10–15 $\mu m^2\ min^{-1}$ (Fig. 1d, f, g, and Supplementary Video 2). During the initial 2 min of constriction, weak constriction pulses were often interrupted by periods of cell stretching. However, at 2 min, constriction pulses increased in magnitude and cell shape seemed to be stabilized between pulses, leading to net constriction (Fig. 1d). These two phases probably correspond to the 'slow/apical flattening' and 'fast/stochastic' phases that have been described previously[13,14]. Overall, cells underwent an average of 3.2 ± 1.2 constriction pulses over 6 min, with an average interval of 82.8 ± 48 s between pulses (mean $\pm$ s.d., $n = 40$ cells, 126 pulses). Constriction pulses were mostly asynchronous between adjacent cells (Fig. 1h and Supplementary Video 3). As a consequence, cell apices between constrictions seemed to be pulled by their constricting neighbours. Thus, apical constriction occurs by means of pulses of rapid constriction interrupted by pauses during which cells must stabilize their constricted state before reinitiating constriction.

To determine how myosin might generate force during pulsed constrictions, we simultaneously imaged myosin and cell dynamics by using myosin regulatory light chain (*spaghetti squash*, or *squ*) fused to mCherry (Myosin–mCherry) and Spider–GFP. Discrete myosin spots and fibres present on the apical cortex formed a network that extended across the tissue (Fig. 2a and Supplementary Fig. 2a). These myosin structures were dynamic, with apical myosin spots repeatedly increasing in intensity and moving together (at about 40 $nm\ s^{-1}$) to form larger and more intense myosin structures at the medial apical cortex (Fig. 2c, Supplementary Fig. 2b, c, and Supplementary Video 4). This process, which we refer to as myosin coalescence, resulted in bursts of myosin accumulation that were correlated with constriction pulses (Fig. 2b–e and Supplementary Video 5). The peak rate of myosin coalescence preceded the peak constriction rate by 5–10 s, suggesting that myosin coalescence causes apical constriction (Fig. 2e). Between myosin coalescence events, myosin structures, including fibres, remained present on the cortex, possibly maintaining cortical tension between constriction pulses (Fig. 2c). Contrary to the purse-string model, we did not observe significant myosin accumulation at cell–cell junctions. To confirm that constriction involved medial myosin coalescence and not contraction of a circumferential purse-string, we correlated constriction rate with myosin intensity at either the medial or junctional regions of the cell. Apical constriction was correlated more significantly with medial myosin (Fig. 2f), suggesting that, in contrast to the purse-string model, constriction is driven by contractions at the medial apical cortex.

Myosin coalescence resembled contraction of a cortical actin–myosin network[15,16]. Therefore, to determine whether apical constriction is driven by pulsed contractions of the actin–myosin network, we examined the organization of the cortical actin cytoskeleton. In fibroblasts and keratocytes, actin network contraction bundles actin filaments into fibre-like structures[16,17]. Consistent with this expectation was our identification of an actin filament meshwork underlying the apical cortex in which prominent actin–myosin fibres spanning the apical cortex appeared specifically in constricting cells (Fig. 3a and Supplementary Fig. 3a). An actin–myosin network contraction model would predict that myosin coalescence results from myosin spots exerting traction on each other through the cortical actin network. To test whether myosin coalescence requires an intact actin network, we disrupted the actin network with cytochalasin D (CytoD). Disruption of the actin network with CytoD resulted in apical myosin spots that localized together with actin structures and appeared specifically in ventral cells (Supplementary Fig. 3b, c). Myosin spots in

[1]Howard Hughes Medical Institute, [2]Department of Molecular Biology, [3]Lewis-Sigler Institute for Integrative Genomics, and [4]Joseph Henry Laboratories of Physics, Princeton University, Princeton, New Jersey 08544, USA.

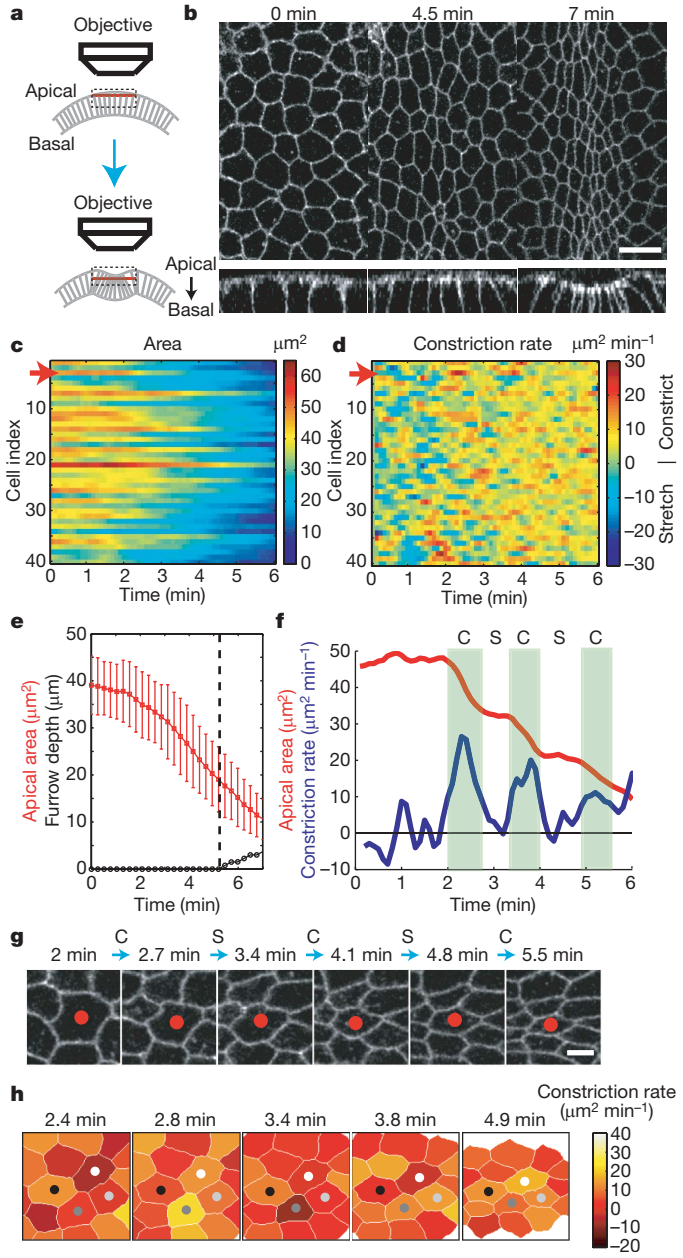


Figure 1 | Apical constriction of ventral furrow cells is pulsed. **a**, Diagram of the imaging approach used to show apical constriction of the ventral furrow cells. We selected tangential *Z*-slices 2 μm below the apical surface (red slices) to show cell outlines. **b**, *Z*-slices (top) and *YZ* cross-sections (bottom) of cell membranes revealed with Spider–GFP. Scale bar, 10 μm. **c**, **d**, Apical areas (**c**) and constriction rates (**d**) for individual cells of a representative embryo. Each row represents data (see colour bars) for an individual cell. **e**, Mean apical area (red) and furrow depth (black). Dotted line indicates when tissue invagination initiates. Error bars indicate s.d. ($n = 41$ cells). **f**, **g**, Quantification (**f**) and time-lapse images (**g**) of the constriction of an individual cell. The red arrows (**c**, **d**) and red dots (**g**) mark the cell that is quantified in **f**. C, contraction. S, stabilization. Scale bar, 4 μm. (**h**) Pulsed constriction is asynchronous in neighbouring cells. Constriction rate is colour-coded (see colour bar) and mapped onto the corresponding cells in images at different time points.

CytoD-injected embryos showed more rapid movement than those in control-injected embryos, suggesting that apical myosin spots in untreated embryos are constrained by the cortical actin network (Supplementary Fig. 3d). Although myosin movement was uninhibited in CytoD-treated embryos, myosin spots failed to coalesce and cells failed to constrict (Fig. 3b and Supplementary Fig. 3e). Because myosin coalescence requires an intact actin network, we propose that pulses of myosin coalescence represent contractions of the actin–myosin network.

Because actin–myosin contractions occurred at the medial apical cortex, it was unclear how the actin–myosin network was coupled to adherens junctions. We therefore imaged E-Cadherin–GFP and Myosin–mCherry to examine the relationship between myosin and adherens junctions. Before apical constriction, adherens junctions are present about 4 μm below the apical cortex[18]. As apical constriction initiated, these subapical adherens junctions gradually disappeared and adherens junctions simultaneously appeared apically at the same level as myosin[8,19]. This apical redistribution of adherens junctions occurred at specific sites along cell edges (midway between vertices; Supplementary Fig. 3f). As apical constriction initiated, these sites bent inwards. This bending depended on the presence of an intact actin network, which is consistent with contraction of the actin–myosin network generating force to pull junctions (Supplementary Fig. 3f). Indeed, myosin spots undergoing coalescence were observed to lead adherens junctions as they transiently bent inwards (Fig. 3c). Thus, pulsed contraction of the actin–myosin network at the medial cortex seems to pull the cell surface inwards at discrete adherens junction sites, resulting in apical constriction.

The transcription factors Twist and Snail regulate the apical constriction of ventral furrow cells[20–23]. Snail is a transcriptional repressor whose target or targets are currently unknown, whereas Twist enhances *snail* expression and activates the expression of *fog* and *t48*, which are thought to activate the Rho1 GTPase and promote myosin contractility[8,10,19,21,24]. To examine the mechanism of pulsed apical constriction further, we tested how Twist and Snail regulate myosin dynamics. In contrast to wild-type ventral cells, in which myosin was concentrated on the apical cortex (Fig. 2a), *twist* and *snail* mutants accumulated myosin predominantly at cell junctions, similarly to lateral cells (Fig. 4a). These ventral cells failed to constrict productively, which supported our cortical actin–myosin network contraction model, rather than the purse-string model, for apical constriction. *twist* and *snail* mutants differentially affected the coalescence of the minimal myosin that did localize to the apical cortex. Although myosin coalescence was inhibited in *snail* mutants, it still occurred in *twist* mutants, as did pulsed constrictions (Fig. 4a and Supplementary Video 6). This difference was also observed when Snail or Twist activity was knocked down by RNA-mediated interference (referred to as *snailRNAi* or *twistRNAi*) (Supplementary Fig. 4a and Supplementary Video 7). However, the magnitude of constriction pulses in *twistRNAi* embryos was greater than that of *twist* mutant embryos, suggesting that the low level of Twist activity present in *twistRNAi* embryos enhances contraction efficiency by activating the expression of *snail* or other transcriptional targets. Myosin coalescence was inhibited in *snail twist* double mutants, demonstrating that the pulsed constrictions in *twist* mutants required *snail* expression (Fig. 4a and Supplementary Video 6). Thus, the expression of *snail*, not *twist*, initiates the actin–myosin network contractions that power constriction pulses.

Net apical constriction was inhibited in both *snailRNAi* and *twistRNAi* embryos (Supplementary Fig. 4b). We therefore wondered why the pulsed contractions that we observed in *twistRNAi* embryos failed to constrict cells. Using Spider–GFP to visualize cell outlines, we found that although constriction pulses were inhibited in *snailRNAi* embryos, constriction pulses still occurred in *twistRNAi* embryos (Fig. 4b, c, Supplementary Fig. 4c and Supplementary Video 8). However, the constricted state of cells in *twistRNAi* embryos was not stabilized between pulses, resulting in fluctuations in apical area with little net constriction (Fig. 4b, c). This stabilization defect was not due to lower *snail* activity, because these fluctuations continued when *snail* expression was driven independently of *twist* by using the P[*sna*] transgene (Fig. 4b)[20]. Although the frequency and magnitude of constriction pulses in such embryos were similar to those in control embryos, stretching events were significantly higher in *twistRNAi*; P[*sna*] embryos, suggesting a defect in maintaining cortical tension

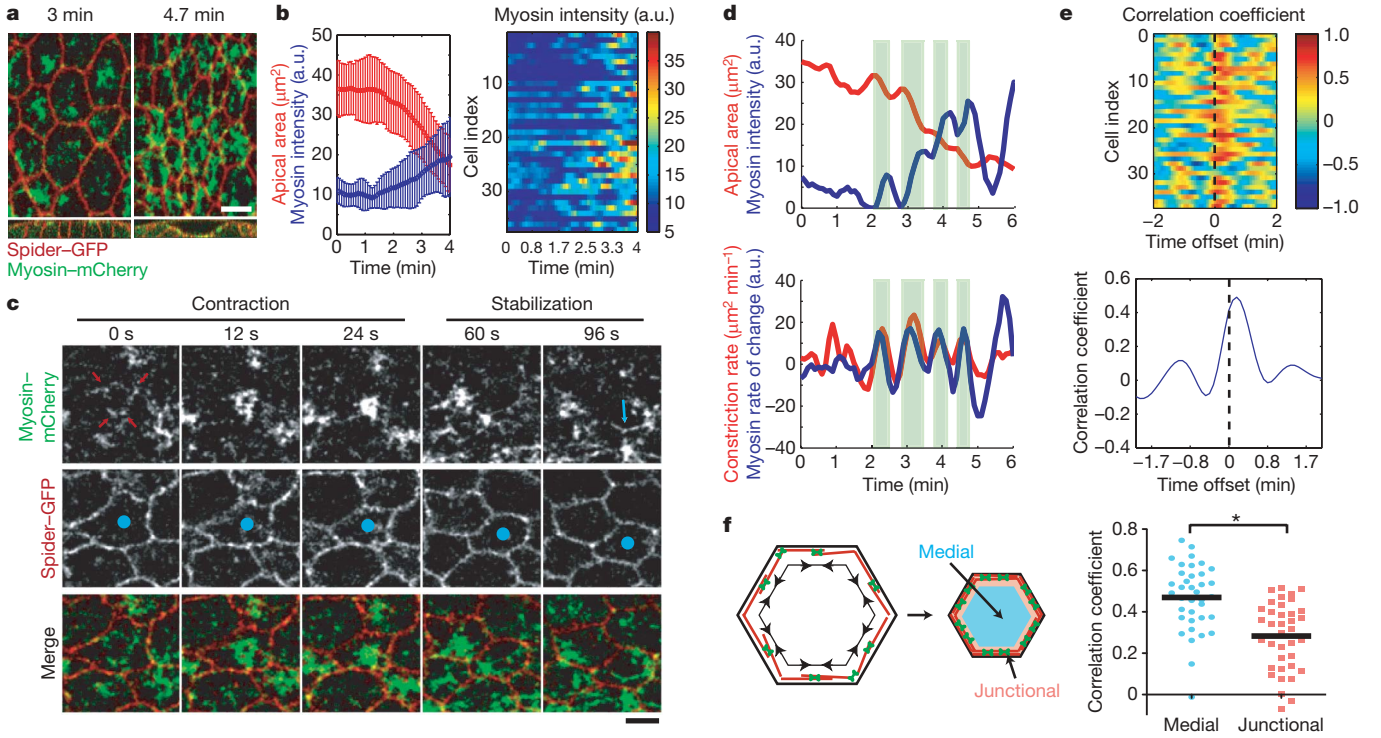


Figure 2 | Constriction pulses are correlated with myosin coalescence. **a**, Merged images of Myosin–mCherry (*Z*-projection, 5 μm depth, green) and Spider–GFP (individual *Z*-slice 2 μm below the apical cortex, red). *YZ* cross-sections at lower magnification to illustrate furrow progression are shown at the bottom. **b**, Mean apical area and myosin intensity (left) and myosin intensity for individual cells (right) for a representative embryo. Error bars indicate s.d. ($n = 37$ cells). **c**, Single channel and merged time-lapse images of Myosin–mCherry (green) and Spider–GFP (red). Red arrows indicate spots that will coalesce. Blue arrow indicates myosin fibre that appears between contractions. **d**, Plots of apical area and myosin intensity against time (top) and constriction rate and rate of change of myosin intensity against time (bottom) for an individual cell. **e**, Plot of correlation between constriction rate and myosin intensity rate of change for individual cells (top) and averaged ($n = 37$ cells, bottom) against time offset. Correlation coefficients were calculated for various time offsets by temporally shifting the data sets relative to each other. Dotted lines indicate zero offset. Note that the maximum correlation occurs when myosin rate is shifted 5–10 s later in time; myosin coalescence therefore slightly precedes constriction rate. **f**, Constriction rate is more highly correlated with medial myosin than with junctional myosin. The diagram (left) illustrates the purse-string model for contraction in which we expect actin and myosin to become concentrated in the junctional region on constriction. Data points (right) represent correlation coefficients for individual cells, and the black bar indicates the mean ($n = 37$ cells). Asterisk, the difference between the means is statistically significant ($P < 0.0001$). Scale bars, 4 μm.

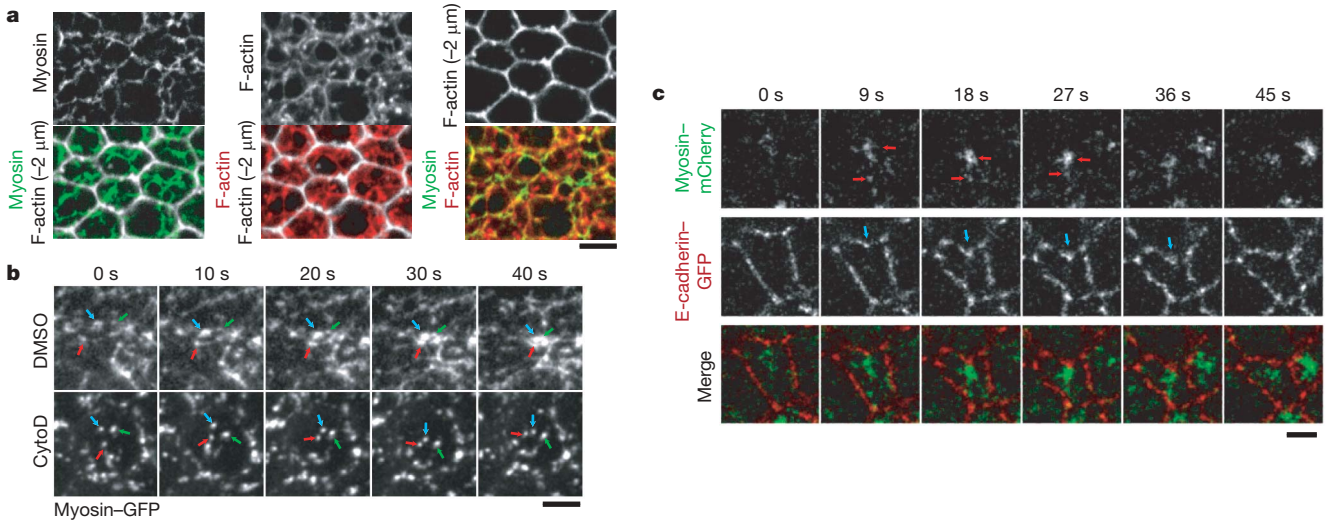


Figure 3 | Pulsed myosin coalescence and adherens junction bending require an actin–myosin network. **a**, Cortical myosin (green), cortical F-actin (red), and F-actin 2 μm below the apical cortex (white, to illustrate cell shape) were revealed in fixed embryos. **b**, Time-lapse images of Myosin–GFP in control-injected (DMSO) and CytoD-injected embryos. Arrows indicate individual myosin spots. Note that myosin spots move, but do not coalesce, in CytoD-treated embryos. **c**, Single-channel and merged time-lapse images of Myosin–mCherry (green) and E-Cadherin–GFP (red). Red arrows indicate myosin coalescence. Blue arrows indicate the site where adherens junctions bend inwards beneath a myosin spot. Scale bars, 4 μm.

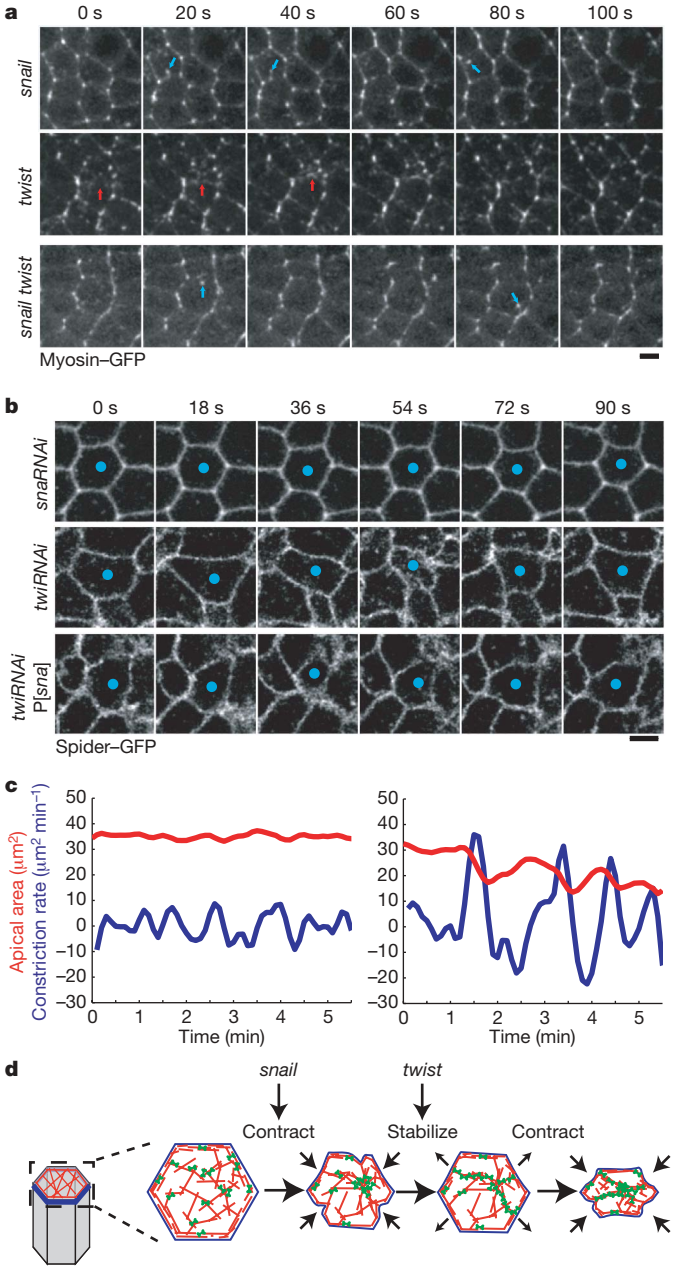


Figure 4 | Snail and Twist function at distinct phases of pulsed constriction. **a**, Time-lapse images of Myosin–GFP *Z*-projections. Blue arrows indicate myosin spots that do not efficiently coalesce in *snail* mutants. Red arrows indicate myosin coalescence in *twist* mutants that seems to pull cell junctions. At least one coalescence event that pulled cell junctions occurred over a 6-min period for 53% of cells in the *twist* mutant, in contrast with 4% of cells in *snail* and *snail twist* mutants ($n = 60$ cells, three embryos per mutant). Scale bar, 4 μm. **b**, Time-lapse images of Spider–GFP in *snailRNAi* or *twistRNAi* embryos. P[*sna*] indicates *twist*-independent *snail* expression. Scale bar, 4 μm. **c**, Quantification of apical area (red) and constriction rate (blue) for individual cells in *snailRNAi* (left) and *twistRNAi* (right) embryos. **d**, Ratchet model of apical constriction. Myosin (green) contracts an apical actin network (red) that is coupled to adherens junctions (blue) driving constriction. Contractions are pulsed, interrupted by a phase in which the constricted state of the cell is stabilized.

(Supplementary Fig. 4d). This defect might result from a failure to establish a dense actin meshwork, because both *twist* mutants and *twistRNAi* embryos had a more loosely arranged apical meshwork of actin spots and fibres than constricting wild-type cells did (Supplementary Fig. 4e). *twist* expression therefore stabilizes the constricted state of cells between pulsed contractions.

Thus, we propose a 'ratchet' model for apical constriction, in which phases of actin–myosin network contraction and stabilization are repeated to constrict the cell apex incrementally (Fig. 4d). In contrast to the purse-string model, we find that apical constriction is correlated with pulses of actin–myosin network contraction that occur on the apical cortex. Pulsed cortical contractions could allow dynamic rearrangements of the actin network to optimize force generation as cells change shape. Because contractions are asynchronous, cells must resist pulling forces from adjacent cells between contractions. A cortical actin–myosin meshwork seems to provide the cortical tension necessary to stabilize apical cell shape and promote net constriction. The transcription factors Snail and Twist are critical for the contraction and stabilization phases of constriction, respectively. Thus, Snail and Twist activities are temporally coordinated to drive productive apical constriction. Despite the dynamic nature of the contractions in individual cells, the behaviour of the system at the tissue level is continuous, in a similar manner to convergent extension in *Xenopus*[25]. Pulsed contraction may therefore represent a conserved cellular mechanism that drives precise tissue-level behaviour.

METHODS SUMMARY

Image acquisition and analysis. Two-colour imaging was performed at 22–25 °C with a Leica SP5 confocal microscope, a 63×/1.3 numerical aperture glycerine-immersion objective, an argon ion laser and a 561-nm diode laser. Spider–GFP images represent confocal slices 2 μm below the apical cortex, whereas myosin images represent maximum-intensity *Z*-projections of an apical section 5 μm in depth. Image stacks for Spider–GFP movies were acquired every 6 s, and image stacks for two-colour movies were acquired every 5 s. Using MATLAB (MathWorks) we developed methods to track cells and measure apical area and myosin intensity. Data points were smoothed with a Gaussian smoothing filter with $\sigma = 15$–18 s (three time points). Myosin intensity was measured from maximum-intensity *Z*-projections as the sum intensity of all pixels in a cell. Mean myosin intensity was calculated for junctional and medial pools of myosin by creating masks that selected regions less than 0.3 μm or more than 0.3 μm from the cell boundary, respectively.

Embryo fixation and staining. Heat fixation and staining with anti-myosin heavy chain (MHC) antibody did not preserve the normal organization of apical myosin observed in live *squ–GFP*[26], *squ–mCherry* (Myosin–mCherry), and *GFP-zipper* (GFP–MHC)[27] embryos. We therefore used endogenous GFP fluorescence to reveal myosin. *squ–GFP* embryos were fixed with 10% paraformaldehyde/heptane for 20 min, devitellinized manually, stained with Alexa-568 phalloidin (Invitrogen) to reveal actin, and mounted in AquaPolymount (Poysciences, Inc.).

Drug/RNAi injection. CytoD was injected laterally at mid–late cellularization with 0.5 mg ml^{-1} CytoD in 10% dimethylsulphoxide (DMSO; Calbiochem). Double-stranded RNAs against *snail* and *twist* (2 mg ml^{-1}) were injected laterally into freshly laid eggs that were incubated 2.5–3.0 h before gastrulation was imaged.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 19 June; accepted 7 October 2008.
Published online 23 November 2008.

1. Lecuit, T. & Lenne, P. F. Cell surface mechanics and the control of cell shape, tissue patterns and morphogenesis. *Nature Rev. Mol. Cell Biol.* **8**, 633–644 (2007).
2. Leptin, M. Gastrulation movements: the logic and the nuts and bolts. *Dev. Cell* **8**, 305–320 (2005).
3. Alberts, B. *et al. Molecular Biology of the Cell* 5th edn (Garland Science, 2008).
4. Baker, P. C. & Schroeder, T. E. Cytoplasmic filaments and morphogenetic movement in the amphibian neural tube. *Dev. Biol.* **15**, 432–450 (1967).
5. Burnside, B. Microtubules and microfilaments in newt neuralation. *Dev. Biol.* **26**, 416–441 (1971).
6. Hildebrand, J. D. Shroom regulates epithelial cell shape via the apical positioning of an actomyosin network. *J. Cell Sci.* **118**, 5191–5203 (2005).
7. Karfunkel, P. The activity of microtubules and microfilaments in neurulation in the chick. *J. Exp. Zool.* **181**, 289–301 (1972).
8. Dawes-Hoang, R. E. *et al. folded gastrulation*, cell shape change and the control of myosin localization. *Development* **132**, 4165–4178 (2005).
9. Fox, D. T. & Peifer, M. Abelson kinase (Abl) and RhoGEF2 regulate actin organization during cell constriction in *Drosophila*. *Development* **134**, 567–578 (2007).

10. Nikolaidou, K. K. & Barrett, K. A. Rho GTPase signaling pathway is used reiteratively in epithelial folding and potentially selects the outcome of Rho activation. *Curr. Biol.* **14,** 1822–1826 (2004).
11. Young, P. E., Pesacreta, T. C. & Kiehart, D. P. Dynamic changes in the distribution of cytoplasmic myosin during *Drosophila* embryogenesis. *Development* **111,** 1–14 (1991).
12. Morin, X., Daneman, R., Zavortink, M. & Chia, W. A protein trap strategy to detect GFP-tagged proteins expressed from their endogenous loci in *Drosophila*. *Proc. Natl Acad. Sci. USA* **98,** 15050–15055 (2001).
13. Oda, H. & Tsukita, S. Real-time imaging of cell–cell adherens junctions reveals that *Drosophila* mesoderm invagination begins with two phases of apical constriction of cells. *J. Cell Sci.* **114,** 493–501 (2001).
14. Sweeton, D., Parks, S., Costa, M. & Wieschaus, E. Gastrulation in *Drosophila*: the formation of the ventral furrow and posterior midgut invaginations. *Development* **112,** 775–789 (1991).
15. Vavylonis, D., Wu, J. Q., Hao, S., O'Shaughnessy, B. & Pollard, T. D. Assembly mechanism of the contractile ring for cytokinesis by fission yeast. *Science* **319,** 97–100 (2008).
16. Verkhovsky, A. B., Svitkina, T. M. & Borisy, G. G. Myosin II filament assemblies in the active lamella of fibroblasts: their morphogenesis and role in the formation of actin filament bundles. *J. Cell Biol.* **131,** 989–1002 (1995).
17. Svitkina, T. M., Verkhovsky, A. B., McQuade, K. M. & Borisy, G. G. Analysis of the actin–myosin II system in fish epidermal keratocytes: mechanism of cell body translocation. *J. Cell Biol.* **139,** 397–415 (1997).
18. Muller, H. A. & Wieschaus, E. *armadillo*, *bazooka*, and *stardust* are critical for early stages in formation of the zonula adherens and maintenance of the polarized blastoderm epithelium in *Drosophila*. *J. Cell Biol.* **134,** 149–163 (1996).
19. Kolsch, V., Seher, T., Fernandez-Ballester, G. J., Serrano, L. & Leptin, M. Control of *Drosophila* gastrulation by apical localization of adherens junctions and RhoGEF2. *Science* **315,** 384–386 (2007).
20. Ip, Y. T., Maggert, K. & Levine, M. Uncoupling gastrulation and mesoderm differentiation in the *Drosophila* embryo. *EMBO J.* **13,** 5826–5834 (1994).
21. Leptin, M. *twist* and *snail* as positive and negative regulators during *Drosophila* mesoderm development. *Genes Dev.* **5,** 1568–1576 (1991).
22. Leptin, M. & Grunewald, B. Cell shape changes during gastrulation in *Drosophila*. *Development* **110,** 73–84 (1990).
23. Seher, T. C., Narasimha, M., Vogelsang, E. & Leptin, M. Analysis and reconstitution of the genetic cascade controlling early mesoderm morphogenesis in the *Drosophila* embryo. *Mech. Dev.* **124,** 167–179 (2007).
24. Costa, M., Wilson, E. T. & Wieschaus, E. A putative cell signal encoded by the *folded gastrulation* gene coordinates cell shape changes during *Drosophila* gastrulation. *Cell* **76,** 1075–1089 (1994).
25. Keller, R., Shook, D. & Skoglund, P. The forces that shape embryos: physical aspects of convergent extension by cell intercalation. *Phys. Biol.* **5,** 15007 (2008).
26. Royou, A., Sullivan, W. & Karess, R. Cortical recruitment of nonmuscle myosin II in early syncytial *Drosophila* embryos: its role in nuclear axial expansion and its regulation by Cdc2 activity. *J. Cell Biol.* **158,** 127–137 (2002).
27. Franke, J. D., Montague, R. A. & Kiehart, D. P. Nonmuscle myosin II generates forces that transmit tension and drive contraction in multiple tissues during dorsal closure. *Curr. Biol.* **15,** 2208–2221 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. Kiehart and R. Karess for providing flies; J. Goodhouse for assisting with microscopy; and S. De Renzis, X. Lu, T. Schupbach, A. Sokac, F. Ulrich and Y.-C. Wang for helpful comments on the manuscript. This work is supported by grant PF-06-143-01-DDC from the American Cancer Society to A.C.M., National Institutes of Health/National Institute of General Medical Sciences grant P50 GM071508 to M.K., and by National Institute of Child Health and Human Development grant 5R37HD15587 to E.F.W. E.F.W. is an investigator of the Howard Hughes Medical Institute.

Author Contributions Biological reagents and fly stocks were made by A.C.M. and E.F.W., and experiments were performed by A.C.M. Image analysis methods were developed by M.K. and the live-imaging data were analysed by A.C.M. and M.K. The first draft of the manuscript was written by A.C.M. All authors participated in discussion of the data and in producing the final version of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to E.F.W. (efw@princeton.edu).

METHODS

Fly stocks and genetics. Fluorescent fusion protein stocks Spider–GFP (III) (*95-1*)[12], Myosin–GFP (II or III) (*squ*^{AX3}; *squ–GFP*, a gift from R. Karess)[26], E-Cadherin–GFP (II) (*ubi-DE-cad–GFP*)[13], and Moesin–GFP (III)[28] are described in the indicated references. Myosin–mCherry (III) (*squ–mCherry*^{A11}) was recombined with Spider–GFP or Moesin–GFP to generate Myosin–mCherry Spider–GFP/TM3 and Myosin–mCherry Moesin–GFP. Homozygous Myosin–mCherry Spider–GFP flies could not be maintained as a stock. Therefore, Myosin–mCherry Spider–GFP/Myosin–mCherry flies were used for two-colour imaging. The dynamics of Myosin–mCherry and Spider–GFP observed in the two-colour strain are indistinguishable from the dynamics of Myosin–GFP and Spider–GFP alone. This suggests that the behaviour of myosin we observe with two-colour imaging reflects normal myosin dynamics. E-Cadherin–GFP Myosin–mCherry flies used for imaging were *ubi-DE-cad–GFP shg*^{R69}/CyO; *squ–mCherry*^{M1}.

Strategies were used to reveal both homozygous and hemizygous *twist* and *snail* mutants. First, the *squ–GFP* transgene was jumped onto the CyO balancer, generating CyO–Myosin–GFP. Chromosomes containing *sna*^{IIG05}, *twi*^{ey53} and *sna*^{IIG05} *twi*^{ey53} were marked with Df(2L)dpp[s7-dp35] 21F1–3;22F1–2 (*halo*) to allow homozygous embryos to be distinguished from their heterozygous siblings. We then rebalanced *twist* and *snail* mutants to obtain *halo twi*^{ey53}/CyO–Myosin–GFP, *halo sna*^{IIG05}/CyO–Myosin–GFP, and *halo sna*^{IIG05} *twi*^{ey53}/CyO–Myosin–GFP. Homozygous mutants were selected by identifying the *halo* mutant phenotype. Myosin accumulation in ventral cells was delayed in *twist* and *snail* mutants, occurring after cephalic furrow initiation, in contrast to wild-type embryos, in which myosin appears before the cephalic furrow is observed. The results presented are representative from nine *halo sna*^{IIG05} movies, six *halo twi*^{ey53} movies and three *halo sna*^{IIG05} *twi*^{ey53} movies.

Alternatively, a compound second chromosome stock homozygous for Myosin–GFP (III) (*C(2)v*; Myosin–GFP) was used. In the *C(2)v* stock, the right arms (containing *twist*) and the left arms (containing *halo* and *snail*) of chromosome 2 assort independently. For *snail* mutants, *C(2)v*; Myosin–GFP was crossed to *halo sna*^{IIG05}*/CyO* and halo embryos were identified. For *twist* mutants, *C(2)v*; Myosin–GFP was crossed to *halo twi*^{ey53} */CyO* and one-third of the non-halo progeny were *twist* mutants. Both strategies identified the distinct phenotypes for *snail* and *twist* mutants described in Fig. 4a.

To provide *twist*-independent *snail* expression we used the P[*sna*$_c$] (referred to as P[*sna*]) transgene, which contains the *snail* cDNA downstream of two copies of the proximal element of the *twist* promoter[20]. One zygotic copy of P[*sna*$_c$] was sufficient to induce invagination in a *snail* mutant background, indicating that it provides enough activity to rescue *snail*-dependent contraction.

Construction of Myosin–mCherry. Myosin regulatory light chain (*spaghetti squash*, or *squ*), including its native promoter, was tagged at the carboxy terminus with mCherry. A 2-kilobase genomic fragment containing the *squ* open reading frame (ORF) and promoter was inserted into the *Kpn*I and *Sal*I sites of pBluescript. The *squ* 3′ untranslated region (800 base pairs) was then cloned into the downstream *Bam*HI and *Xba*I sites. The sequence for mCherry, including a short linker region, was cloned into the *Cla*I and *Bam*HI sites in between the *squ* ORF and 3′ untranslated region. The 3.5-kilobase *Kpn*I/*Xba*I fragment containing Myosin–mCherry was then cloned into pCasPer4 and sent to BestGene Inc. to make transgenic flies. Myosin–mCherry complemented the null *squ*^{AX3} allele, demonstrating that it is functional.

Time-lapse image acquisition. Egg collections were performed in plastic cups covered with apple-juice plates. Flies were allowed to lay eggs for 2–4 h at 25 °C before the plate was removed and embryos undergoing cellularization were collected. Embryos were dechorionated with 50% bleach, washed with water, and then mounted on a slide with embryo glue (Scotch tape resuspended in heptane), with the ventral side facing upwards. A chamber was made with two no. 1.5 coverslips as spacers and was filled with Halocarbon 27 oil for imaging. Embryos were not compressed. Mesoderm invagination occurred with a timeframe similar to that deduced from fixed embryos, and embryos imaged under these conditions could subsequenty hatch, demonstrating that our imaging conditions had minimal impact on development.

Single-colour images of Spider–GFP and two-colour images of Myosin–mCherry Spider–GFP were obtained with a Leica SP5 confocal microscope, a 63×/1.3 numerical aperture glycerine-immersion objective, an argon ion laser and a 561-nm diode laser. Images were acquired with a pinhole setting of two Airy units. For simultaneous two-colour images, we set the excitation bandpass to 495–550 nm to detect GFP, and to 578–650 nm to detect mCherry. There was minimal bleedthrough between the two channels. Images were acquired at a resolution of 141 nm per pixel. Myosin–GFP images were obtained with a PerkinElmer Ultraview spinning disk confocal microscope, a 60×/1.2 numerical aperture water-immersion objective, an argon/krypton laser and an Orca ER 4 charge-coupled device camera (Hamamatsu).

Image processing and analysis. The images presented were processed with ImageJ (http://rsb.info.nih.gov/ij/) and Photoshop CS (Adobe Systems, Inc.). Unless otherwise noted, Spider–GFP images represent a single confocal slice 2 μm below the apical cortex, whereas Myosin–GFP, Myosin–mCherry and Moesin–GFP images represent maximum-intensity *Z*-projections 5 μm in depth. A Gaussian smoothing filter with a radius of one pixel was used to reduce noise in published images.

To show apical cell shape, we manually selected *Z*-slices at a depth that was about 2 μm below the apical surface and the apical myosin. Cell outlines at 2 μm depth were very similar in dimensions to more apical cell outlines; however, they were easier to reveal and lacked membrane irregularities (namely in Fig. 3c) that complicated image analysis. Shifts in *Z*-position did not result in discontinuities in the data (Fig. 1e), and the pulsed behaviour that we describe was also observed when a single *Z*-slice was used.

We developed methods to measure apical area, constriction rate and myosin intensity with MATLAB (MathWorks). Raw images were bandpass-filtered with effective cutoff wavelengths of 1.4 μm (low pass) and 17.9 μm (high pass). After thresholding, a series of morphological operations were applied to reduce the width of the membranes to one pixel and to remove spurious background labelling. Examples of extracted cell outlines are shown in Supplementary Fig. 1. Indexed cells were automatically tracked on the basis of distances between cell centroids at subsequent time points. Cell properties were measured at each time point. We manually removed cells with errors in the segmentation to ensure that all cells in the data set were correctly identified. Data for apical area and myosin were smoothed with a Gaussian smoothing filter ($\sigma = 15$–18 s, three time points), and constriction rates and myosin rates of change were calculated from the smoothed data. Unless otherwise stated, a constriction pulse was defined as an event in which the constriction rate exceeded one standard deviation above the mean (more than 10.8 $\mu m^2\,min^{-1}$). To measure myosin intensity, we first clipped intensity values below two standard deviations above the mean to separate myosin structures from unselective background labelling. Myosin intensity was then measured from maximum-intensity *Z*-projections (two highest values averaged) as the sum intensity of all pixels in a given cell. The correlation between constriction rate and the myosin intensity rate of change was determined by calculating the correlation coefficients between these two data sets for individual cells using the time interval from 1.7 min to 4.8 min. To examine the time dependence of this correlation, the data for constriction rate and myosin rate of change were shifted in time relative to one another.

Furrow depth (Fig. 1e) was calculated by measuring the distance between vitelline membrane and apical cortex in *YZ* cross-sections.

Myosin–GFP spot velocity was measured using the Manual Tracking plugin for ImageJ. We tracked individual spots or other distinct myosin structures for the duration of their lifetime and averaged the three highest velocities to calculate maximum velocity. For wild-type movies, we calculated the velocity of myosin structures coalescing. We chose coalescence events that occurred early or occurred near the midline to minimize the effects of tissue movement during invagination.

Drug injection. Embryos were dechorionated in 50% bleach for 2 min, washed with water, mounted on the edge of a glass slide (ventral side upward) with embryo glue, and desiccated for 4–8 min. A 3:1 mixture of halocarbon 700:halocarbon 27 was used for injection. Embryos were injected laterally at mid–late cellularization (furrow canals had passed the base of the nuclei) with about 1% egg volume of control or drug solution. We injected 0.5 mg ml^{-1} CytoD (Calbiochem) in 10% DMSO.

RNAi. Primers for double-stranded (ds)RNA were designed with E-RNAi (http://www.dkfz.de/signaling2/e-rnai/)[29]. Primers included the sequence of the T7 promoter (5′-TAATACGACTCACTATAGGG-3′) followed by the following recognition sequences: Twi01-F, 5′-GCCAAGCAAGATCACCAAAT-3′; Twi01-R: 5′-GACCTCGTTGCTGGGTATGT-3′; Twi02-F, 5′-GGAGCTGCAGAACAATGTGA-3′; Twi02-R, 5′-TGCTGTTGCTGGTGGATTAG-3′; Sna01-F, 5′-CGGAACCGAAACGTGACTAT-3′; Sna01-R, 5′-GCGGTAGTTTTTGGCATGAT-3′; Sna02-F, 5′-ATCATGCCAAAAACTACCGC-3′; Sna02-R, 5′-AGCGACATCCTGGAGAAAGA-3′; Control-F, 5′-GAATGCTATGGGAGGCGATA-3′; Control-R, 5′-TCAGCTTAGGCTCTGGGTGT-3′.

Primer pairs were used to amplify a PCR product from genomic DNA. PCR products were used directly in a transcription reaction with T7 polymerase with the MEGAscript transcription kit (Ambion). Annealing was performed by adding 10 mM EDTA, 0.1% SDS and 0.1 M NaCl to the reaction and incubating this mixture in a water bath heated to above 90 °C, which was allowed to cool for several hours. The dsRNA was purified by extraction with phenol/chloroform and resuspended in injection buffer (5 mM KCl, 0.1 mM sodium phosphate, pH 7.0). We injected a 2 mg ml^{-1} solution of *snail* or *twist* dsRNA into the embryo. Identical results were obtained with either dsRNA fragment used to knock down *twist* or *snail*. Sna01 and Twi01 were used for all the experiments in

the manuscript, with the exception of Supplementary Fig. 4e, in which Twi02 was used. Control primers amplified a portion of an unknown ORF, CG3651, which has no function during ventral furrow formation. Injections were performed as described for the drug injections, except that egg collections were performed after 30 min to inject embryos at the earliest possible stage. To provide *twist*-independent *snail* expression we crossed Spider–GFP virgin females to P[sna_c][20] males and collected embryos for *twist* dsRNA injection.

Statistics. Statistical significance between means was determined with an unpaired *t*-test. *P* values were calculated with Prism 5 (Graphpad Software, Inc.).

28. Edwards, K. A., Demsky, M., Montague, R. A., Weymouth, N. & Kiehart, D. P. GFP–moesin illuminates actin cytoskeleton dynamics in living tissue and demonstrates cell shape changes during morphogenesis in *Drosophila*. *Dev. Biol.* **191,** 103–117 (1997).
29. Arziman, Z., Horn, T. & Boutros, M. E-RNAi: a web application to design optimized RNAi constructs. *Nucleic Acids Res.* **33,** W582–W588 (2005).

CORRIGENDUM

doi:10.1038/nature07735

Major gradients in putatively nitrifying and non-nitrifying Archaea in the deep North Atlantic

Hélène Agogué, Maaike Brink, Julie Dinasquet & Gerhard J. Herndl

Nature 456, 788–791 (2008)

The *y* axis in Fig. 3 of this Letter was incorrectly labelled 'CO_2 fixation (m^{-3} day^{-1})'. It should read 'CO_2 fixation (μmol C m^{-3} day^{-1})'.

naturejobs

THE CAREERS MAGAZINE FOR SCIENTISTS

This issue introduces the 2009 class of *Naturejobs* Postdoc Journal keepers. As usual, the competition for our four spots was intense. This year's crop was chosen from 60 applicants from 20 countries, working in a variety of fields — from marine-resource management to behavioural neuroscience. The winners demonstrated wit, a knack for writing, intriguing backgrounds and a keen insight into the challenges and dilemmas facing postdocs of all backgrounds.

Three of the journal keepers have more than research to keep them busy, as parenthood has introduced the familiar challenges of a work–life balance. Journal keeper Julia Boughner, a British-born postdoc in evolutionary developmental biology at the University of Calgary in Canada, describes her predicament as "the race to secure a tenure-track position before family and other personal demands drive me out of my dream of academia". Boughner models the mechanisms of human craniofacial variation in a morphometrics lab.

Joanne Isaac, studying the impact of climate change on tropical systems and species at James Cook University in Queensland, Australia, is also juggling a postdoc, motherhood and a relationship. Isaac delved into the life history of brushtail possums as a graduate student. Bryan Venters, a postdoc at Pennsylvania State University at University Park, is a new father weighing up the merits of industry versus academia for his future. Venters studies gene regulatory mechanisms as a postdoc; he mapped genome-wide locations of yeast transcription before that.

Sam Walcott described a different but increasingly common set of challenges: how to plan one's career and training as an interdisciplinary scientist. Trained in theoretical and applied mathematics, Walcott is a theoretical biophysics postdoc at Johns Hopkins University, Baltimore, where he focuses on molecular biomechanical models related to muscles. He wonders if a fledgling interdisciplinary scientist, by focusing on just one discipline during his postdoc, adversely affects his career.

I hope readers will look forward to tracking our journal keepers' progress as they pursue their own career aims. And I offer my thanks to all those who applied.

Gene Russo is editor of *Naturejobs*.

CONTACTS
Editor: Gene Russo
Assistant editor: Karen Kaplan
e-mail: naturejobseditor@naturedc.com

European Head Office, London
The Macmillan Building,
4 Crinan Street, London N1 9XW, UK
Tel: +44 (0) 20 7843 4961
Fax: +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

European Sales Manager:
Dan Churchward (4966)
e-mail: d.churchward@nature.com
Assistant European Manager:
Nils Moeller (4953)
***Nature*events:**
Ghizlaine Ababou (+44 (0) 20 7014 4015)
e-mail: g.ababou@nature.com
Southwest UK/RoW:
Alexander Ranken (4944)

Northeast UK/Ireland:
Matthew Ward (+44 (0) 20 7014 4059)
France/Switzerland/Belgium:
Muriel Lestringuez (4994)
Scandinavia/Spain/Portugal/Italy:
Evelina Rubio-Hakansson (4973)
North Germany/The Netherlands/Eastern Europe: Kerstin Vincze (4970)
South Germany/Austria:
Hildi Rowland (+44 (0) 20 7014 4084)
Advertising Production Manager:
Stephen Russell
To send materials use London address above.
Tel: +44 (0) 20 7843 4816
Fax: +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com
***Naturejobs* web development:** Tom Hancock
***Naturejobs* online production:** Dennis Chu

US Head Office, New York
75 Varick Street, 9th Floor,
New York, NY 10013-1917
Tel: +1 800 989 7718
Fax: +1 800 989 7103
e-mail: naturejobs@natureny.com
US Sales Manager: Ken Finnegan

India
Vikas Chawla (+91 1242881057)
e-mail: v.chawla@nature.com

Japan Head Office, Tokyo
Chiyoda Building, 2-37 Ichigayatamachi,
Shinjuku-ku, Tokyo 162-0843
Tel: +81 3 3267 8751
Fax: +81 3 3267 8746

Asia-Pacific Sales Manager:
Ayako Watanabe (+81 3 3267 8765)
e-mail: a.watanabe@natureasia.com
Business Development Manager, Greater China/Singapore:
Gloria To (+852 2811 7191)
e-mail: g.to@natureasia.com

NATURE|Vol 457|22 January 2009

MOVERS

Linda Birnbaum, director, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina

2008–present: Senior toxicologist, National Center for Environmental Assessment, Research Triangle Park, North Carolina
1989–2008: Director, Experimental Toxicology Division, National Health and Environmental Effects Research Laboratory, Research Triangle Park, North Carolina

After 16 months without a director, the US National Institute of Environmental Health Sciences (NIEHS) in Research Triangle Park, North Carolina, has a new head. Toxicologist Linda Birnbaum, who has spent much of her career at the Environmental Protection Agency (EPA), took the helm at the institute this month.

Birnbaum started off as a biologist at the University of Rochester in New York, and was soon attracted to the burgeoning fields of molecular biology and molecular genetics. Her PhD at the University of Illinois in Urbana saw her map the ribosomal RNA genes of *Escherichia coli*. After a postdoc at the University of Massachusetts in Amherst and a stint at Kirkland College in Clinton, New York, Birnbaum took a post at the Masonic Medical Research Laboratory in Utica, New York, where she studied ageing. Here began her career-defining work in toxicology.

In studying how altered metabolism could affect ageing, Birnbaum modulated metabolism in rats using dioxins and polychlorinated biphenyls. She looked at how these chemicals break down in the human body and for the next 10 years explored the molecular effects of dioxins, analysing the relative toxicity of related chemicals, and designing long-term bioassays to assess cancer risk. "If anybody had told me that 30 years later I'd still be working with that family of chemicals, I wouldn't have believed them," she says.

Taking over as head of the EPA's Environmental Toxicology Division in 1989, Birnbaum expanded its ranks to an all-time high of 90 full-time employees and dozens of students and postdocs. Despite budget cuts, she found ways to continue toxicology studies by partnering with other federal agencies or academia. She spearheaded some of the first work documenting the mechanism of action of endocrine-disrupting chemicals, such as brominated flame retardants.

Kenneth Ramos, president of the Society of Toxicology, hopes that Birnbaum will help the NIEHS, and the maturing field of environmental health, better define itself. "With an internationally recognized toxicologist as its leader, the institute can now focus its efforts and have an impact on disease causes," he says. "Linda has the expertise and conviction to inspire and grow the institute on many levels."

Birnbaum is encouraged by the incoming administration of Barack Obama's stated commitment to science, health and the environment. "I plan to create a holistic approach that can deal with the biggies, from complex mixtures of toxic chemicals to climate change," she says. ■

Virginia Gewin

NETWORKS & SUPPORT

Research assistants join a union

Research assistants at the Research Foundation of the State University of New York (SUNY) in Stony Brook have decided to unionize — the latest development in ongoing unionization battles at US universities. Nearly all who voted last month to join the Communication Workers of America (CWA) are working in science, says Matthew Engel, a Stony Brook research assistant who campaigned for union representation. Frustrated by issues such as fees and job insecurity, they are seeking benefits comparable to those received by teaching assistants.

Research assistants are graduates temporarily doing academic research at a college, university or non-university institution. They usually work on projects supervised by full-time academics who administer the funds that provide their salaries. Teaching assistants generally receive fixed pay from the university.

In 2004, the federal National Labor Relations Board ruled that research assistants are students, not employees, and so could not be represented by a union. But a 2007 board ruled that those at the SUNY Research Foundation in Albany, Buffalo and Syracuse were fundamentally employees.

Teaching assistants are already represented by the CWA. "We found out that they had negotiated some benefits, and we thought it was a good idea," says Engel, a doctoral student in Stony Brook's biomedical engineering department. According to the union website, its local branch now represents more than 4,000 research and teaching assistants throughout the SUNY system. Engel hopes to negotiate reduced fees and improved pay and health benefits.

SUNY research assistant Luigi Longobardi emphasizes the importance of job security for non-US research assistants whose visa status depends on funding. "If we're going to be without funding, we should have a fair amount of time to find another adviser or alternative sources of funding," he says.

According to the Coalition of Graduate Employee Unions, some 30 unions cover graduate employees on more than 60 campuses.

"The Research Foundation is committed to following all laws and regulations related to collective bargaining," a foundation spokeswoman says. The university has raised no objections to the vote.

Engel says the research assistants are selecting a negotiating team and expect to sign a contract with the Research Foundation soon. ■

Karen Kaplan

POSTDOC JOURNAL

A pregnant pause

What is the science behind short-term memory loss? As the sleep-deprived mum of a six-month-old boy, I'd enjoy knowing the technical aspects of my hamstrung brain. I study craniofacial development and evolution with the aim of understanding the molecular mechanisms underlying coordinated morphological change in the teeth and jaws, mostly in primates.

I've had frequent memory lapses since returning to the lab in December after half a year of leave. Finicky bench work was challenging enough when my pre-offspring focus was at its sharpest. I will have to compensate for my sideways mind by finding tactics that enable me to be productive while managing the baby's demands on my time. I am my own work–life balance guinea pig.

The next 12 months will be about finding my balance as a partner and a parent, regaining my stride as a scientist, and achieving my career goals. I aim to be a tenure-track assistant professor with my own lab and minions. (What's a bona fide scientist without minions?) Will I cut it? Or will I cut out? Private industry may offer me a more liveable (and lucrative) work–life equilibrium. I must thoroughly research how my skills could best be applied in an enjoyable non-scientific career. And I must talk to people who have successfully made this leap. This is my quandary and my journey as a postdoc. ■

Julia Boughner is a postdoc in evolutionary developmental biology at the University of Calgary, Canada.

Replacement

Welcome back.

Shelly Li

You stand over her grave, tears splattering onto the 100-year-old dirt. It's been ages since you saw her in the flesh, since you watched that crooked smile spread from one end of her face to the other.

God, she was so beautiful, so flawless … and you're so drunk, you almost think about dropping to your knees, digging for her. But instead you clutch the whiskey bottle closer to your heart and tell her, like you always do: "Happy birthday", and walk away.

The next year you are back, and the next. You know what you want to happen, but after so many years, the pain never fades, and the memory of her has never left your mind for a single second.

Then one afternoon you show up, high off your ass and ready to collapse in front of the tombstone, when you see someone else standing in your spot.

"Hey," you say, stumbling over. "What are you doing?"

The man turns to you. "You must be the poor bastard she left behind." He turns off the PRIVATE reading in his brain chip, giving you his name and information.

Owen Powers, Brain-Chip Transferor

You understand and don't have to read further. You look up at the man, study him for a while. He looks to be about 20, 25 years old, but in the year 2230, who can tell the difference? He may well be, like you, rounding the year 200.

"So," you say. "It's been so long, I was afraid that you people didn't even exist."

"Well I'm here now, aren't I?"

You pause after that, think for a minute before saying: "And you're here to offer me your services?" You've spent too many years on this godforsaken planet to display too much emotion, only to be stabbed by lies, over and over again.

But the words that come out of Powers's mouth next make your heart beat to an irregular tune again, bring life back into your old, empty soul. "Yes. We will return your wife to you," he says.

A little smile touches your lips, but you don't let it travel far before stomping it back down again. "It's been more than 50 years."

"Yes, it has, and we're sorry that it's taken this long to move down the wait list to you. Our scientists had some technical issues to resolve on the new models, but we think that you will be pleased with the progress we have made."

This time a wind of joy flies in and slams into your chest, lights you all the way up.

JACEY

"Okay," you say, "then how does this work?"

"We will pull your wife's brain-chip records, put them inside a body and ship it to you. You won't get to choose the body we put her in, but we'll try to make an android that looks close enough."

You shrug. It was never the outer appearance that was the most beautiful, but the angel on the inside. The only question left is: "Will she be the exact same person?"

"The brain-chip recorder stops as soon as the heart does. Trust me, it'll be like she never left."

The doorbell rings two weeks later, and you stand up, shuffle to the door.

Each step seems heavier than the one before, as each heartbeat becomes harder and harder to slow down.

Love, warmth, any feeling you thought you'd lost forever when she left is back again, pumping through your body like liquid fire.

Your finger hesitates for a second on the OPEN DOOR button, but you finally unfreeze yourself and pound on it.

The door slides open …

Your whole body stops, and your brain spins into orbit.

It's her. It's really her.

You reach out, pull her into your arms and take a deep breath …

But then you realize that something's wrong.

She used to smell like pomegranate and raindrops. Now all that's left is the scent of nothingness.

Her hair, she had a cowlick in the back, where a patch of rebel hair always stuck up. This woman standing before you, however — her hair is annoyingly perfect.

Her outer shell feels like the skin of a human, but you know that she never will be. All she's done is taint your memory of *her*.

"Come in," you say with a smile on your face, taking her hand and leading her into the house.

You walk with a tightening inside you to the drawer, pull out your gun. Before she can even say a word, you pump three bullets into her chest. She falls to the floor, crackling with electricity, emitting a screen of a smoke. Amid the fire starting in your living room, you squeeze a final shot, and she stops moving completely.

As you stand over her body, tears start to slide down your face once again. "I'm sorry," you say, "but an android can never replace her."

You scoff at yourself and look around the big, empty house. This was a stupid, stupid idea. Love and loneliness blinded you, drove you to accept this offer.

But what's done is done.

Now the curtain of tricks has parted, and you see the truth.

The sealing hollowness of it. ■

Shelly Li is a 15-year-old who aspires to finish high school and go to college, perhaps get in a few rounds of golf in between. She can be found at www.shelly-li.com.